National Quality Forum

Moderator: Sheila Crawford
July 2, 2014
1:00 p.m. ET

Lauralei Dorian: Thank you. Good afternoon, everyone. This is Lauralei Dorian on the (Encraft) team here. Thank you for joining the (inaudible) Family Center (inaudible) workgroup call.

Just a note to please keep your lines on mute. It just helps with the sound. And also a reminder that this call is open to members of the public, and we will be having a public comment period at the end of the call.

We just wanted before we got started to make sure we know who's on the call. I can see on the webinar I believe we have Sharon Cross.

Sharon Cross: Yes, I'm here.

Lauralei Dorian: Great. Hi, Sharon. And Esther Neuwirth.

Esther Neuwirth: I'm here.

Lauralei Dorian: Great. And are there any other – do we have (Leigh) or Ann or Liz Mort on the call yet?

(Leigh Partridge): This is (Leigh). I am on the phone, and I'm working on trying to get on the web.

Lauralei Dorian: OK, great. Thanks, (Leigh).

(Samuel Burner): And I'm (Samuel Burner). I'm on the call.

Lauralei Dorian: Oh, hi, (Samuel).

Eric Coleman: Hi, Lauralei. Eric Coleman and (Truce Joyce) are on the call, and we're also working on getting on the web link.

Lauralei Dorian: OK, great. Hi, Eric. And do we have any other measure developers on the call?

Lucille Schacht: This is Lucille Schacht at the NRI.

Lauralei Dorian: Hi, Lucille.

Lucille Schacht: Hi.

(Joanne Scorasteen): This is (Joanne Scorasteen) from (west stat) for the (caps) measure.

Lauralei Dorian: Great. And do we have Anne Monroe or Liz Mort yet? OK.

So, thank you very much for joining the call today, and thanks especially to the committee members who submitted surveys and spent time reviewing these measures. We know they can be quite daunting and complicated, and so we appreciate you taking the time to do that. And we're going to talk through the measures – you know, each of the three measures today, but we're also happy to take questions about or feedback about the process in general.

As I said, we understand that it has been quite time consuming. These measures are a little bit different than our standard quality measures. So we can talk through a bit of the differences later on in the call.

Actually, I'll talk through to them now. We wanted to remind you that NQF endorses – does not endorse surveys themselves. These measures, as you know, are based on survey data, but we – (but) the individual person level and (inaudible) performance measures that are at a facility level in order to determine performance at that level. So, we expect (testing) at both of those levels, at the both the data element level for the survey itself but also at the performance score for the facility level. So, we'll ask the developers as we go through each of the measures to describe that testing at both of those levels. Did you have any (inaudible)?

And then I think the process for the call today for our first measure, we're going – which is (0228) which is the (CPM-3) measure, we're going to go through each of the survey questions individually and then probably for the next two measures depending on how long the first measure takes, we'll sort of – we won't go through each question individually, but we'll take more – we'll go through each of the criteria at a higher level. And so we wanted before we got into individual measures, we just wanted to ask whether you had any feedback about the process, how it went for you or questions that might pertain across the board to all of the measures that aren't measure specific.

Karen Pace: And as we go along – this is Karen Pace. You know, feel free – we can also revisit this later if you have some thoughts to share with us. The other thing just so you know for today we really want to use this as an opportunity for – as we go through the criteria for you to become more familiar with the criteria to – we can help find things in the form that relate to the criteria for you to raise questions or issues, and as we heard, the developers on the call so we can directly ask them questions if we can't find something or have a question about the submission or their measures. So I guess with that I will just stop again and see if anyone has any questions – overarching questions or comments. Otherwise, we'll get into our first measure. (All right, Lauralei.)

Lauralei Dorian: So, we're going to begin with measure, as I said, 0228, which is the three-item care transition measure, the (CPM-3) from the University of Colorado. And this is a hospital-level measure of performance, which reports the average patient reported quality of preparation for self-care response among adult patients discharged from general acute care hospitals within the past 30 days.

And so if you are connected to the webinar, you can see on your screen the first question that we asked you, which relates to evidence to support the measure focus. And we asked if measuring a health outcome or a PRO which of course is the relationship between the PRO and at least one healthcare action which can include a structure, a process, intervention or service, identified and supported by the stated rationale.

Karen Pace: So, we'll go to the evidence form in the materials. And just to clarify again as Lauralei was reading the question, we asked you to respond to just to clarify that our evidence for outcomes and PROs (or) a type of outcomes as experienced with care really just requires a rationale, and it's definitely great if the rationale is supported by evidence, but we're not expecting the same kind of systematic review and grading of a body of evidence that we would for a process performance measure or an intermediate clinical outcome.

So, basically, we asked the developer to first identify what kinds of structures or processes of care would affect the outcome – in this case, the experience with being prepared for transition – and then to give a rationale. So, if you're in the measures submission form, you would have seen what (Nadine) has up on the screen now on the webinar. And we'll just stop there and see if anyone has any questions.

I think from the standpoint of the committee, the (inaudible) your point (inaudible) thing that you should be thinking about is does this make sense? Is it logical? And then, you know, does the rationale really support those relationships that they mentioned? So, I'll stop there and see if there are any questions or comments.

OK. So, I think, you know, this – and just so everyone knows, this is a sufficient amount of information. We don't – aren't asking for pages and pages. This is the kind of information we're looking for.

One of the things that I will point out, though, to everybody, because we've got some other submissions we'll be looking at, this particular measure submission is really one performance measure. Some of the other ones we'll see and have seen the caps measures, each submission may have eight to 15 individual performance measures. So from the standpoint that we're kind of looking at information for one measure, I think it makes it a little bit easier certainly to find things and to have a clear picture of what's going on.

So, Lauralei, you want to – and (Nadine) go to 1B the question.

Lauralei Dorian: So, the next question, 1B, asks about performance gap. So we're asking does the performance data provided demonstrate a gap in care. It can be variability

of cross entities or overall (inaudible) optimal performance to warrant a national performance measure. And we also asked if the performance data provided demonstrate disparities for certain population (inaudible). So we can bring the performance gap section up on the submissions form.

(Samuel Burner): Can I – I just want to ask. So, the provider, the developer, has to demonstrate that. That either there's disparity among some population or there's a gap that has to be demonstrated as rationale for the measure.

Karen Pace: Yes, and when – would everyone please identify themselves. I know we're at a disadvantage with everybody being on a disconnected phone line here. We can't see you.

But just to answer your question is yes. All the information, all the responses come directly from the developer, and it is up to them to provide information that shows that they're meeting the NQF criteria.

(Samuel Burner): OK. I raised my hand on there to show who I am.

Karen Pace: Oh, thank you. Thanks, (Sam). I sometimes have difficulty working in more than one media at a time.

Lauralei Dorian: And I see – do we have Liz Mort on the call now?

Liz Mort: Yes, you do.

Lauralei Dorian: OK great. Hi, Liz.

Liz Mort: Hi.

Karen Pace: So, you see here that the developer was able to provide some information about the distribution of scores. And basically, this came from (me). And the other thing that I'll mention for those of you who are familiar or perhaps were on the call the other day is that this CTM measure is also being incorporated into the hospital (caps). So, we'll see it in both places, but this is the individual measure, meaning it could be used outside of the (caps) system.

So, any questions or comments about the data that were provided? And, you know, just another example. When we have the (caps), you'll be seeing national data for most of those, unless they're new and have just had limited testing. In this case, the data are from Maine, because that's what the developer had access to. So when you're looking at performance gap, you also have to be kind of thinking about how representative it is. So even if there wasn't a big performance gap, for example, for Maine, you would ask yourself the question is that representative across the country. So…

Any questions?

(Samuel Burner): What is that – OK. Somebody's jumping around a lot on the slide. It's a little hard to figure out what you're showing us.

Karen Pace: Right.

(Samuel Burner): You're showing us some (descriptive) statistics here. OK.

Karen Pace: Yes. Let's…

Lauralei Dorian: Where did you jump to? I don't know where you're at.

(Inaudible)

Karen Pace: OK.

Female: There is 1B2.

Karen Pace: Right. Let's just look at 1B2 and see – yes, just leave it there for a second. So here you see that information by question and really what we want in this section is the computed performance score. So that's additional information but not required and not necessarily what you would be basing your decision on. So the – then if – below that, they have information that's coded by hospitals. So there's example you'll see hospital AA, AB, AC, et cetera. And so let's just ask the developer maybe.

Eric, could you explain what these columns are? Because if this is an individual hospital, what is their computed performance score?

Eric Coleman: Sure, I'd be happy to. And your timing was great, as my computer tried to sleep the second you asked the question. (So) we have paper backup here.

So, we have had a terrific partnership with the state Maine who really was ahead of the rest of the country in thinking about the importance of really a person level measure in thinking about performance reporting. And so Maine is, you know, certainly not a large state by standards – other standards – but they did worry about the (inaudible) get participation across all acute care hospitals. And as Karen indicated the far left column and is just a code for the hospital. As we read across this over a particular time window how many individual consumers or patients provided a response to the survey.

And then that gives us the hospital level means of the hospital lower confidence limit mean as well as -- sorry, your columns are a little different than mine -- the lower confidence level that probably should be the upper confidence limit. I apologize I think that's a typo on our part. And then a comparison to the benchmark which is the statewide mean and aggregate of all the participating hospitals.

Karen Pace: OK. All right.

(Samuel Burner): You're showing a upper and lower confidence limit? Because I -- it's not labeled quite -- I'm a little confused because of the labeling. So the -- like on the top number there 85.75 is the mean and then 83.86 is the lower confidence interval of the mean and the upper limit is 87.95?

Eric Coleman: That is correct and we apologize for that typo. You're absolutely correct.

(Samuel Burner): OK.

Karen Pace: OK, any other questions or comments about this? So, we will -- so you can kind of look at, if you go down the and look at the -- each hospitals mean which is what their score on this particular measure would be and compare that to the state average. You can get a sense of some of the spread and variability across those hospitals.

OK, and then the other thing that we ask you to think about and we asked the developers to provide is any information about potential disparities or differences in performance related to for example, ethnic or racial groups or income status, et cetera. And doesn't mean that there has to be but we asked for that because performance gap could relate to either variability across facilities or that there's a gap in care related to certain population subgroups.

So, let's see, do you have any (inaudible) there? That would be in 1B4, right, all right. So, they didn't have any data regarding that from this measure as specified but did provide some information in 1B5 of what might be known from the literature.

Lauralei Dorian: Are there any comments or questions about disparities?

Then we might move to the next …

(Samuel Burner): It looks like this is just a three question that were used, doesn't look like they had any issues with different ethnic groups. I image the state of Maine there's not a lot of Spanish speaking people so I guess they didn't -- not look at it in Spanish speaking. I don't see that listed. I doubt they have many people that speak Spanish.

Karen Pace: OK, so you want to go on to 1C?

Lauralei Dorian: Yes, we'll move on to 1C which is high priority. We asked specifically for a PRO-PM how it was determined that the target population values the (PRO) and find it meaningful. So we asked the developer to provide some information on working with families and patients for example on how they determined what questions would be asked on the survey.

So we can go to that section on the measure information form. Scroll down just a little.

Karen Pace: 1C5, so a couple of things, obviously this addresses a, you know, a population leaving hospitals but probably more importantly for these (clients) of patient reported measure that, you know, 1C5 asks about the involvement of patients

and consumers and really identifying is it measuring something of importance to them or valued by them.

So, they reported on their focus groups. Any questions or comments about that?

(Samuel Burner): And this was -- that looks like they established some content validity with the focus groups.

Lauralei Dorian: Right.

(Samuel Burner): So, looks good.

Karen Pace: OK. All right, Lauralei.

Lauralei Dorian: So that's going to move on to specifications now. First we asked in terms of (inaudible) specifications for reliability which specifications, if any, are unclear? So that could be the data element definitions, the codes, the steps in the calculation, the case mix adjustment.

And for a (Pro PM), is it clear which questions and items and possible response options are used for the performance measure? Are there sampling instructions? And do you have any concerns that the likelihood of this measure can't be -- that the measure couldn't be consistently implemented?

So those are a few things asked on 2B1. And also 2B1, which is specifications consistent with the evidence, so are the specifications consistent with what the target population values and finds meaningful? So we'll open that up, there are a few things in both of those questions, for any thoughts, comments or?

Karen Pace: So if you're -- we'll go to (S4) the numerator or S, yes, (S4) and you'll see that we talk about the numerators the hospital level, some of the CTM3 scores for eligible sample patients. When you get down into the details (S6) they give the specific questions that go into the CTM measure and they're response.

So the question would be is, let's see, what they do is they sum the number that's associated with the response option for the hospital and probably we can go down to the scoring algorithm, (S18). So here they give the specifics about how the data are aggregated to come up with the facility level score.

(Samuel Burner): OK.

Karen Pace: And then some sampling and survey …

(Samuel Burner): Could you go a little lower on that page?

Karen Pace: OK.

(Samuel Burner): After step three I guess.

Karen Pace: There we go. OK.

(Leigh Partridge): Karen, this is (Leigh Partridge).

Karen Pace: Yes.

(Leigh Partridge): My question's really directed to Dr. Coleman. I was a little confused on this scoring method because as I understand it this measure is going to be incorporated in (HCAPS). And the measures that we're reviewing that are the other (HCAPS) measures that we're reviewing in general use the what -- the (CAPS) team refers to as top box scores.

Eric Coleman: So, thank you, (Leigh), for that question and, you know, first of all just a big thanks to Karen and Lauralei for helping us along in this process. It is obviously a bit of a dance to purpose continuing use of the CTM as an individual measure (inaudible) at the same time have the opportunity for CTM to be incorporated into the (HCAPS).

And I think one of our guiding principles is to try to avoid creating confusion in the field that people who are using the CTM3 whether as a separate standalone measure or whether incorporate into (HCAPS) are not having to go -- switch back and forth between specifications and calculations and things.

So, all along we've been working with the (HCAPS) folks to make sure that the way this measure functions and the characteristics are aligned with how they're using it as as well. We do not have our own individual top box scores to report but we have been collaborating with the (HCAPS) folks and we did include some of that experience as they've been doing a much more ambitious level of national testing then we've had resources to. And we've shared those as part of this application.

(Leigh Partridge): But I think that the question is they will be computed differently, is that correct? Yours is based on the sum of the scores and converting to this 0 to 100 -- and then the mean and then converting to a 0 to 100 score versus (HCAPS) is going to use the top box scoring method?

Eric Coleman: Yes, you know, as I have had a chance to review (HCAPS) data I've seen them report both using the conversion scale but I'd be happy to do a little research on that and follow up with Karen and Lauralei.

Karen Pace: Right and it's something, you know, we looked at the (HCAPS) but we didn't focus on this but I think (Leigh's) got a good point but it will be one of the things that, you know, that the committee may want to sort of explore with the developers if they are being scored differently, you know, to understand that and whether that's necessary, so.

Eric Coleman: We're flexible on this end and we're certainly open to that kind of -- those kinds of suggestions.

(Leigh Partridge): Yes, I think our real question's going to be is there any way of knowing whether the different scoring systems are going to be radically different in results? Is the hospital going to look different if you use …

Eric Coleman: No.

(Leigh Partridge): The CTM alone versus the (HCAPS) scores?

Eric Coleman: No, that's an important question and from what I understand about how the top box scores works, you know, what's underneath the hood is still

comparable. It's really how it's actually being reported. So we'll work with the team on that.

Karen Pace: Right, but the concern would be, you know, NQF is endorsing the performance measure which includes how the score's going to be computed for each -- for the accountable entity. And, you know, the point (Leigh's) making is the big issue, you know, if they get scored, you know, if they get two different scores how do you interpret that? So, we'll have more discussion about that I'm sure.

Any other questions about the specifications and scoring? So, one of the things and I guess we'll move on to testing but that is a question though kind of move through testing is because we expect the testing to be on the measure at the specified.

So, we'll look at the testing but and, Eric, we'll be asking you to describe that in just a minute, but that will be one of the questions is the testing that you presented is it based on the scoring that you put in these specifications or you also mentioned dome information from (CMF) that's good as a measure test.

Oh, wait a minute, sorry I jumped ahead. I'll let Lauralei guide us through the questions first.

Lauralei Dorian: So, the first question is about reliability. So we asked what level was tested and again we expect the survey itself on the performance measure to be tested. (Those) reliability testing with an adequate scope, meaning the number of entities and patients to generalize for wide spread implementation and with an appropriate method and describe how the results either do or do not demonstrate sufficient reliability.

So we'll ask Eric to talk us through how reliability was conducted at (both) levels.

Eric Coleman: Thank you, Lauralei, I was just making sure I've got the right stuff in front of me, OK.

Karen Pace: Yes, let's go down the (2As).

Eric Coleman: Column 2A2?

Karen Pace: Yes, right.

Eric Coleman: So, you know, the character (addition) measure has really been fortunate to have been around for a bit having gone through the endorsement process twice and so we do have along the way a cumulated different experiences in terms of testing of the measure and with (Leigh's) comments well taken.

So, the we have been collaborating with the (Bill Erman) and the (HCAPS) -- (CMS HCAPS) folks, and have results around the reliability based on their experience, which is with a fairly significantly larger sample. But we also, in our own testing, have reliability to report as well. And I don't know how much detail Karen and (Lorelei) were looking for here, but I'd be happy to sort of walk people through the individual pieces here.

(Leigh Partridge): Right. Would you just maybe briefly explain how the reliability of the performance score was tested, what your (exploits) is, so, you know, is this based on the computed hospital score? And I guess the question is what score, based on HCAHP's methodology or your methodology?

Eric Coleman: Right. So, the answer is both. In a way we do have sort of a hybrid of testing. The most recent testing is HCAHP, using their methodology. And then we also have – using our own resources when we went out to the field to perform, to evaluate the performance of the measure. And those are – that was using the calculations that we just reviewed.

You know, as we have looked at this and, you know, we do have card-carrying psychometricians, methodology people, I have to confess that is not my card-carrying, but we have gone through using the more traditional ways of evaluating a measure, such as confirmatory factor analysis (going back) to also being one was of reporting that. We've also had an opportunity to use some other methodology that refers to something called (routes) and compare the two.

And essentially under (2a2.3), performance measure score testing, we see, going back to (all the internal) systems be 0.8, (experiment brown) being 0.84. So, you know, in terms of translating this into language for people who don't spend all their days doing measurement, this puts the measure into, certainly, a category of acceptability. Certainly there could be higher levels of reliability but this, I think, by most measurement experts would say that the measure does perform in a manner that we have some confidence.

(Leigh Partridge): Thank you. And so the chron box (alpha), are those – so are these results that you're presenting, are they based on the ...

Eric Coleman: They're based the calculations that we've provided, yes.

(Leigh Partridge): OK. But HCAHPS doesn't compute it the same way. That's what I'm – I just want to be clear.

Eric Coleman: Right. And so, we do report the way that HCAHPS has done their calculations as well.

(Leigh Partridge): So, OK. I guess I'm missing it, because there's only one number here. So, are these numbers based on – the hospital level of reliability. Is that based on how you calculate it or how HCAHPS calculates it?

Eric Coleman: The one in (2a2.3) are our calculations. And I think, (Trish), my colleague, is pulling up the HCAHPS stuff. Sorry. Oh, these are HCAHPS. OK. OK. Sorry.

So, sorry, we're just comparing here. So, (Trish), who was instrumental in helping us get prepared and fill out the forms, has pointed out that the ones in (2a2.3) are the HCAHP results.

(Leigh Partridge): OK.

Eric Coleman: And then the one underneath it says critical data elements testing are the results that we did using our approach.

(Leigh Partridge): OK. All right.

Karen Pace: Any questions or comments from the committee?

OK. Why don't we look at the question regarding the (slide) on validity testing?

Lorelei Dorian: (And to me ask), again with testing conducted at both levels with an adequate scope to generalized (life study) compensation, and we asked, did the results demonstrate sufficient validity. The conclusions about quality can be made. We'll bring up the validity section, and ask you again to walk through briefly what your strategy for validity testing was.

Eric Coleman: Yes, OK. So, (inaudible) (QV3). So, Lorelei, would you like us to sort of walk through the steps we've taken, or just go to (2b2.3) which are where the results?

Lorelei Dorian: I guess, explain the steps that were taken, briefly, and then go over the results, if you would.

Eric Coleman: OK. And once again, you know, we really have sort of a parallel path here. As indicated earlier, we had this opportunity to partner with the HCAHPS developers and work with them along the way as they did a much more ambitious national sample of over 3 million people. But we've also – in getting at the question of validity, have done a series of focus groups, some of which we made reference to earlier, comparing different constructs to the measure, to make that they – that we're being true to what consumers said was most important to them.

It's one thing for – to have heard them. It's another thing to have successfully translated that information. And the validity step gives us an opportunity to really confirm if we got it right. In other words, the constructs that they said were important, can we bring in some sort of external standard to bounce that up against, to make sure that we're on the right path.

So we actually went through a series of studies with the initial focus groups, as well as going out and evaluating whether – for example, if essentially we're looking at a performance measure that gauges whether hospitals are doing a good job of preparing patients to go home and take care of themselves,

listening to their preferences, helping manage their medications. And one of the manifestations, if things weren't being done well, would be that there would be a much higher likelihood that they would have to return for additional care.

And so, looking at, for example, readmission rates has been one of the steps that we've used, to use a form of validation of the measure. And we have been able to demonstrate that the way consumers answer those questions does, at some level, is predictive of whether they're going to be readmitted or not.

(2b23) provides some of the specific numerical results of the statistical testing using HCAHPS, so validity in this case was measured using a correlation called (Pierson). And the advantage of working with the HCAHPS team is that they have the ability to not only evaluate the CPM in and of itself, but also how it relates to other of the sub-domains of HCAHPS.

So, in the middle of the page, it talks a bit about whether consumers recommend a hospital, how they judge nurse communication, and nurse communication or any health professional's communication around medication, and be able to do some comparisons within that. Also, when you're doing validity testing, you want to make sure that – this is what we sort of call the Goldilocks phenomenon where, you know, you don't really want too hot or too cold. You sort of want just right.

And if you had a measure that consumers responded to and, say, you know, you ask them, on a scale of one to five, and everybody pretty much gave you a five, you wouldn't really know what to do with that. Because it could be that there was a problem in the way that the individual response categories were conceptualized. In other words, it's sort of called the ceiling effect, where everyone is kind of bunched up at the top, and then you're really not able to draw very many conclusions from that.

And similarly you don't want everybody kind of gathered at the bottom, because it's also hard to know what to make of that. A measure that's sort of doing a good job and determining variation and, since we're at a fairly high stakes here, what's a measure that's used for performance and potentially

associated with payment, we'd like to know with some certainly that we can capture a range of responses and that the measure's going to be sensitive to pick up those changes. And what we've been able to show is that there really isn't a ceiling effect. There certainly is room for hospitals to get better on this measure, and that we would be able to detect that if they were getting better. And so this really gives us equity and fairness and things.

We also looked into whether, you know, the characterization measure scores might vary depending on the gender of the person responding, whether or not certain diseases, like heart failure or chronic lung disease, you know, might have scored differently, and then we would get into concerns that a hospital that treated a lot of people with one condition but not the other might be disadvantaged.

So we looked at this in terms of individual conditions. And then we also asked patients about other sort of related experiences that might track along with the elements that are being explicitly measured.

We've done some work where we also wanted to explore whether or not African-Americans, Hispanic-Americans, rural Americans, might have different scorings, in which case either this would need to be brought to the surface. We were able to find that the performance of the measure across different populations was consistent, that there wasn't some notion that the measure resonates with (inaudible) populations but not others. And ...

(Samuel Burner): Yeah, can I – Can I ask you – I'm sorry, I (inaudible) want to ask a question.

Eric Coleman: Please do.

(Samuel Burner): You just – what you just said, was this – all this testing was for your – this was – was any of it done with people in a Spanish language version, or all in an English language version?

Eric Coleman: Yes. You know, for what we have – well, for the studies that we have led or collaborated on, we've not testing a different language version of the care transition measure. However, we have – the measure has been translated now into 17 languages. And we make those tools available.

(Samuel Burner): OK.

Eric Coleman: But we have not necessarily done the testing ourselves. And in many cases people have come forward representing those communities and said, "We'd really like to translate this." And we've said, "Please, by all means. And please let us know what your experiences are."

But since we didn't have a direct hand in that, we were not entirely comfortable reporting this.

(Samuel Burner): OK.

Eric Coleman: Because our (insurance) data was not as high as with this data.

(Samuel Burner): Because I live in Texas where about a third of our population, many of them speak Spanish, and (done that) as their primary language. And – so, it's available in a translated form that's been – but it hasn't been necessarily validated in a Spanish language form.

Eric Coleman: To our knowledge, it hasn't been very rigorously validated. It's certainly been used in the field. People went through the appropriate steps where you initially translate it and then you back-translate it. And so that level of rigor is there.

(Samuel Burner): It's (inaudible), OK.

Eric Coleman: But we don't know that they've gone on to do some of the steps that NQF would be looking for . . .

(Samuel Burner): OK.

Eric Coleman: ...(inaudible) of application.

(Samuel Burner): Thank you.

Lorelei Dorian: OK, great. Are there any other questions about reliability or validity, sorry?

OK, well, we're going to move on to – I'll read it out loud. Feasibility and usability, right after each other. But, so feasibility asks, "Which of the required data elements are not routinely generated and used during care delivery?" And (inaudible) not available in electronic format? And do you have any concerns about how the data collection strategy can be (put into) operational use?"

And then to move onto usability and use, it asks how the measure is being publicly reported. And for maintenance measures, which this is, which accountability applications is the measure being used for, and how can the results be used to further resolve the (inaudible) quality (inaudible) healthcare? And are there any unintended consequences?

And so we'll open it up to (soft instructions) about these two areas. OK. And are there any, before we move on to the next measure, any last minute questions or thoughts about this measure?

(Leigh Partridge): This is (Leigh). I just have a little question for (Eric), because I know from having listened in to some of the other workgroup calls, that some of our members will probably ask it at – this question.

This is a measure in which you don't allow proxies to respond?

Eric Coleman: You know, (Leigh), you're very perceptive. So, we – you know, if you look at what we've primarily been working on the last three years, it's been a very strong focus on the role of family caregivers. And so we are really a little bit internally conflicted about this, because we do recognize that family members have important insights to share, and in many cases they really are at the forefront of what's going on in an individual's care. In some cases, perhaps even more aware than what the individual consumer or patient is going through.

The internal conflict is that we're trying to, as we discussed earlier, have the (CTM3) always be sort of the (CTM3), and not – under some instances it looks like this, and other circumstances it doesn't.

We have had fairly lengthy conversations with the HCAHPS developers about this same point, trying to see if there's room for them to perhaps reconsider the role of proxies in responding, and I imagine that, you know, into the future, we're going to start to see a shift. And we would be very supportive of that shift. But for now we've been a little bit constrained by trying to, sort of, mirror what HCAHP is doing, and currently is, to my knowledge, does not include proxies.

(Leigh Partridge): Thank you.

Lorelei Dorian: All right. So I think we've on to the next measure which is (0726), the inpatient consumer, or ICS, consumer evaluation and inpatient behavior healthcare services, from (NRI). And this was developed to gather patients' evaluation of their inpatient psychiatric care, and it's composed of six individual measures (organ) names.

So, we're off and we won't go through each individually but, in general, did you have any reflections on the evidence criteria, which asks first, (1a), which is, "Is the relationship between the measured outcome and at least one healthcare action identified and supported by the rationale?" And then (1b) asks about performance gaps, so, "Does the performance data provided demonstrate a gap in care?" And we'll bring each of these sections up as we discuss them.

And then (1c) asks how it was determined to target population values (the PRL) and finds it meaningful. First we'll bring up . . .

(Leigh Partridge): And this is a – this survey has – looks like six performance measures?

Lorelei Dorian: Six, yes.

Karen Pace: Six performance measures. So each one of those should be addressed in the evidence, meaning is – are there – for each one, is there information about what healthcare interventions can affect it? And then when we get to specifications, of course, each one needs to be specified as well as reliability and validity testing needs to be provided for each of those six measures. So, let's look at the – OK. This is the – what they've provided. So I'm not sure.

This (1a2) looks – this looks like the process of giving the information verse – or getting the survey filled out, and computing the measure. Or computing the measures. And then (1a2.1) seems to be some general information about improvement. But I'll just ask the committee that's looked at this if you have any comments or questions, and then we can ask if the developer can provide some more information.

So, do we have someone from the measure developer on the line?

Lucille Schacht: Yes, hi. This is Lucille Schacht at NRI.

(Leigh Partridge): OK. So, what we're looking for here is, based on each of these measures, to identify some of the healthcare interventions and actions that will actually influence the – you know, the experience that you're trying to measure. So, do you have any additional information for that, or any thoughts about that?

Lucille Schacht: Well, the – in other parts of this as well, we do specify information about what each one of these domains is measuring. Concepts like dignity, rights, the environment, empowerment, and how those questions are related to the patient's perception of their experience of the interaction in their care and the – during the inpatient stay.

What's on your screen now is some information about the rights domain. We did talk about each of the domains throughout the application.

Karen Pace: Right. So, we'll come back to this. This – we'll just see if the committee has any questions or wants any more information. Generally we want this for each of the six measures, but we'll move on to looking at performance gaps.

Lorelei Dorian: OK, so, again, for performance gaps it asks whether the information provided demonstrates a gap in care (for one) and national performance measure. We also ask about disparities information. So we've brought performance gap up on the screen.

Is there any questions or comments about performance gap?

Karen Pace: So, maybe we'll just ask Lucille to – I know you've provided this for each measure, but just to kind of put it into perspective, let's look at measure one, and (1b2), the results.

Yes, if you would just explain kind of what this information means.

Lucille Schacht: Sure. There are, as you pointed out there are six different domains. And so what we've provided here is the results from 2008 analysis versus the 2011 analysis, to show change over time as well as the – within each time period what was the means, the range of values. Each indicates that the – each of the domains has some variation across facilities, so that you can measure differences at facilities. And there's been some general improvement over time, over the three years of the studies.

Lorelei Dorian: OK, fine.

Karen Pace: So, I have a ...

(Samuel Burner): Can you – I'm sorry. Can you lower the slide so we can see the other three? You've shown the top three. Can you go down and ...

Karen Pace: Right. Yes. Let me just ask Lucille to clarify first, because we are asking here for distribution of the facility scores, and it looks like you have a note here, percent of patients responding positively.

So is this all patients aggregated together or is this based on actual facility scores, this distribution.

Lucille Schacht: That distribution is the facility scores.

Karen Pace: OK.

Lucille Schacht: And our nomenclature of scoring is the patient scored positively on the items in the domain.

Karen Pace: OK, great thank you.

Lucille Schacht: So that's basically how that is. So this is the range across facilities. The first part of the table will show you the 2008 scores for the six domains. And then, what you see there is the first three on the screen. And then the last three. And then it repeats that information for the 2011 data.

(Samuel Burner): Did you – you're showing aggregated data. Did you look at it by a facility? I mean, this means are for the 36,000 or whatever surveys in total, right?

Lucille Schacht: No. This is the means for the 70 to 80 hospitals in each of the two years.

Karen Pace: Right. And …

Lucille Schacht: So, this is the facility variation.

Female: Right.

Karen Pace: So, it's – I think it says 90 states in psychiatric hospitals in 24 states. So this is the kind of information …

(Samuel Burner): OK. So this is by – OK. So this is …

Karen Pace: Yes. Yes.

OK. And then, you provided also for 2011, and Lucille, you want to just mention briefly what trend you're seeing or no change, or I know …

(Samuel Burner): A lot of your max – your max value seemed higher on this second run.

Karen Pace: Right. And, you know, and that is usually an indication of a concerted improvement effort at the facilities who have been using this over a number of years to actually start impacting their scores. So, you would expect to see some trending up for some of the facilities. And we do see that in a few of the domains.

(Off-mike)

Lauralei Dorian: And we'll move on to disparities.

(Off-mike)

Lauralei Dorian: Like they did provide later from 2008 and 2011. The gender, age, scroll down a little.

(Off-mike)

Lucille, did you want to interpret where these significant differences between these groups?

Lucille Schacht: In general, there – these are the score changes between eight and 11. And, you know, looking at the scores, there are very few differences. We have a huge N, so it's actually easy to get significant differences even though the score may only go from 73 to 74. And that's one of the things that we take into consideration when we're running some of our testing and doing our validations.

So, this section here is actually looking at the changed scores between eight and 11 with any to the different groups. So, by length of stay or long-term status and those kinds of characteristics.

Lauralei Dorian: Great. Thanks. Are there any questions or comments about that?

(Samuel Burner): No.

Lauralei Dorian: OK. Moving on to the top priority which is does the measure address significant health or quality problems and for a PRO-PM what is determined – how is it determined at the target population values, the PRO and finds it meaningful?

Is there any concerns about …

Karen Pace: Maybe let's go down to 1C5 about how they involved patients or what they say about that.

So, it basically talks about the – on traits in reliability and validity. I think the last line says that the former patients were included in original development. So maybe – Lucille, if you want to expand on that a little bit?

Lucille Schacht: Sure. When we initially started this survey back around 2000, we had a CAS Group that involved former patients of hospitals, hospital administrators, quality directors, researchers. And when we did our pilot test, we also provided on the form for patients to provide some feedback about whether these questions hit the target for them, and to, you know, be able to provide additional comment on the survey itself. So we involved patients in the very early stage of development. And at our facilities that use these surveys, they involve patients in the quality initiatives to use the responses for improvement.

Karen Pace: OK. Any questions or comments from our committee members?

(Samuel Burner): I just wanted to ask. Was this being administered by a telephone or how was the response being elicited from the patients?

Lucille Schacht: It is a self-administered paper survey. The patients are given the survey before discharge and they're asked to complete it and leave it at the facility. They can choose to take it with them and mail it back to the facility.

(Samuel Burner): OK.

Lucille Schacht: And some facilities do get mail back, but most surveys are actually done before the person leaves and the paperwork is left there.

(Samuel Burner): That may explain why you have that higher response rate that you mentioned.

(Leigh Partridge): I think it's actually a low response rate. This is (Leigh). Am I right?

Lucille Schacht: Well, we're currently at a 62 percent response rate …

(Leigh Partridge): OK.

Lucille Schacht: … on average …

(Samuel Burner): So, it's pretty good.

Lucille Schacht: … for a self-administered survey.

(Leigh Partridge): No, that's – OK. I think I picked up the wrong data then. I thought it was low.

Lucille Schacht: Yes. No. We started off low and a lot of facilities worked on improving their patient response by helping to show the patients that they were using this information to inform the way that they interacted with patients, and the treatments, and all of that. So, I think that's really driven up the response rate.

(Leigh Partridge): Have you done any – I'm curious – sometimes people are reluctant to turn in a survey and they're leaving because they are afraid that the provider will read it and to somehow or rather take out, you know, be unhappy with them. Have you – did you do any testing before you decided to do it that way to make sure that the patient felt comfortable handing it in the end of their stay?

Lucille Schacht: There's been a lot of testing in the mental health arena around survey distribution and the different impacts of anonymous versus identified surveys, distribution by somebody on your clinical team versus, you know, someone in sort of administrative function. And so, we sort of have that history to go on in the mental health field in terms of how surveys had been distributed.

Most surveys are actually anonymous. So, the patient's name is not on the survey and it clearly states at the survey that their response would not be used in a negative way in relation to their discharge or their planning.

(Leigh Partridge): Thank you.

Lucille Schacht: You're welcome.

Lauralei Dorian: All right. If there are no questions or comments, we can move on to specification.

And so we ask which specifications are unclear. Is it clear which questions, and items, and response options are used to performance measure in other survey or sampling instruction? And do you have any concerns that both likelihood that that this measure can be consistently implemented as well?

And then for 2B1, specifications consistent with the evidence in what ways, if any, of the specifications inconsistent and are they consistent with what the target population values and find it meaningful?

And NQF actually has a – we have a pre-comment period now where we ask numbers of the public to comment on the use of these measures or any – they can submit any comments on maintenance measures. So, we do have a comment. I'm just going to read out loud.

It says, "Each domain is scored as the percentage of clients at the time of discharge or at annual review and respond positively to the domain from the survey for a given month. This is troublesome, should the person be hospitalized multiple times with the first time being satisfied, but the next time not. It would be more beneficial based on most recent hospitalization. This would allow it to be tied to at the time of discharge or at annual review stated in the first sentence under measure description."

So, I'll just ask if the developer has any response to that or if any of the committee members have any other thoughts?

Lucille Schacht: From a developer's perspective, each episode of care should be evaluated by the person which is why it's distributed at the time of discharge. And one episode of care may have gone better than another and patients may respond differently. There may be patients who are not voluntary and that is a piece of information actually in the survey to say that this person was not voluntarily admitted. And we can do stratified reports by that kind of status. And we actually do for our members, do stratified reports.

Karen Pace: So, let's go down to S.4, that's where the specific measures and items are located and then we can look at the scoring. Keep going. OK. Go down with the measure.

Did we get that? OK. I'm sorry, S.6. Let's go down, keep going. Sorry. Thank you. There we go.

On this section of the form as we requested, they identified the performance measure and then the specific items that are used in that performance measure.

And then we can go down to the calculation algorithm, S.18. And this is where the S.18.

And maybe, Lucille, while we're bringing that up, there it is, if you want to just briefly explain the, you know, how it's scored in terms of the performance measure. We've heard the committee already talk about with the CAHPS these, the top box, percentage of patients that have the most positive score. We've heard about average scoring with the CTM. So, could you just explain your scoring?

Lucille Schacht: Sure. As I indicated, the mental health community has been doing surveys for a number of years. I actually with the NRI here worked on an outpatient version of the survey which is also scored on a one to five scale for each item but then domains are scored as an average of a domain. And that sort of just been traditional on the way that we've approached surveys for – every survey you have to decide on a cut point.

We went with the standard that had been used for our consumer surveys as the average score being above that sort of neutral territory, thus the three and half were scored basically on a one to five scale. So three and a half is above neutral as an average score. And that's what we present as a domain score.

Karen Pace: So, when you say positive responses. That would be a four or five on the individual patient's response?

Lucille Schacht: No, it's their average score. So you would add up their scores on each of the four or three or four items in that domain.

Karen Pace: OK.

Lucille Schacht: And then divide it by the number of questions that they have. So they may have a three, four, and five and their average is a four, and so they would be considered to as responded positively.

Karen Pace: And then for the hospital then, you count the number of the patients then that are above the 3.5, is that what you're saying?

Lucille Schacht: Yes. That's how the hospital score is calculated, the percent of people who were above that 3.5 threshold.

Karen Pace: OK.

(Samuel Burner): And so if a person – if a person only answers three of the – if there's four or five responses, they only answered three of them properly that they – it's only if they're less than two then you throw it out it's – that's not calculated as a domain score then?

Lucille Schacht: Right. And we only have three or four items in a domain. And you need to do two items in order to do an average. So, most – half of our domains are three items, the other half are four items. We have very few questions that are actually not answered. And so we throw out very few domain calculations.

(Samuel Burner): OK.

Karen Pace: OK. Any other questions about specifications?

All right, then we will move on to reliability and validity testing. And we'll bring those up in the testing attachment and basically we want reliability testing at both the instrument or the scales that the domain that they referred to, and then for the computed facility score.

So, I'll just ask Lucille if she can give us a very brief synopsis of how they did testing at the scale level and then at the computed hospital score level.

Lucille Schacht: Sure. So the data element reliability was done with a Cronbach's alpha. So it's looking at the correlation among the items. And the – we did by domain. We also did it for age groups. We view a lot of these different tests on a regular basis to sort of re-verify our validity. And then for the performance scores, we basically used ANOVA to look at the hospital-hospital variation. And comparing that to what's called the, you know, the patient-to-patient variation to ensure that the instrument can pick up the differences across hospitals.

So those are the two basic kinds of tests that we did.

Karen Pace: OK. Great. We'll go – let's look at the results in 2A2.3. So, OK, let's go down just here. We're most interested because – NQF is interested in the individual measures. We're most interested in these by the care domain which is how the measures are computed. So, Lucille if you want to just say anything about the Cronbach's alpha? They look …

Lucille Schacht: Well, overall the testing showed strong domains. And also that, you know, if an item was deleted, it would still be a strong domain. That may suggest that we potentially could conclude surveys even if they didn't have the two items. This is something that we're now thinking about. So, all of our testing showed very strong relationships.

Karen Pace: Great. OK. And then we can go down to the results for the hospital level score in terms of reliability – there we go.

These also look very high. Is there a minimum number of responses that are needed to get this high on reliability testing?

Lucille Schacht: I really don't know.

Karen Pace: OK.

Lucille Schacht: The – we have a lot of surveys and what we tend to do is do random samples of parts of that group when we run some test to make sure that our significance is not due to a huge end.

Karen Pace: OK. Right. OK, any questions about the reliability testing from the committee members?

(Samuel Burner): No.

Karen Pace: OK. Let's go on to validity testing then, 2B2. And, Lucille, I'll just ask you again to just briefly describe what you did at the scale level versus the computed performance score.

Lucille Schacht: OK. At the data element level for the scales, we're looking at the Pearson correlation which is looking at the items within the domain to the overall domain average to make sure that every item is really important. And then for

the performance and measures themselves, we looked at – you can scroll down – so.

OK. So, on the performance measures, we were looking at the items in the domain in relation also to an anchor item, our item 28. We considered it to be sort of a globalized measure of satisfaction, so looking at sort of the relationship in a linear regression model with that anchor item and the items in the domain.

Karen Pace: All right. Thank you. And why don't we go down to the results then, 2B2.3? And here, you see that the – these kinds of tables are provided for each of the performance measures. And, Lucille, I'll just ask you to maybe explain …

Lucille Schacht: Sure. You're breaking up a little bit. But these correlations are all significant, indicating that each of the items has a strong relationship to the overall domain score.

Karen Pace: Great. OK.

Lucille Schacht: And the tables are similar for all of the domains.

Karen Pace: All right. And let's move down to the results for the performance measure score validity testing. Keep moving past table six.

OK. And, Lucille, if you want to just walk us through this table.

Lucille Schacht: Yes. This is a correlation between the domain and the variation across the hospitals, assuming a relationship with this anchor item as a measure of that, the strength of the items. And these were all significant. The only one that was actually somewhat low was the outcome of care with the lowest score. And I think that's the one area that was a little troubling, but it is still significant.

Karen Pace: Right. And what is this kind of anchor item, like the global – what is it asking?

Lucille Schacht: The item 28 is – I think it's – if I had a choice of hospitals, I would choose this one.

Karen Pace: OK. And what's the ranking scale or the response scale for that particular item?

Lucille Schacht: All the items are scored the same way. They're on a one to five scale.

Karen Pace: OK. Thank you. OK. Committee members, any questions, comments?

(Samuel Burner): No, it's – it looks pretty good. What – do you have any idea of why you think those outcome of care domain was a lot lower? I mean, it's significantly lower than all the others.

Lucille Schacht: I'm not sure. We haven't had a chance to go in and dig deeper to see if we can determine what this might be related to. It was still significant as a significant relationship even though it has the smallest R-square. So, it is something that we do want to go in and take a look at.

Karen Pace: And maybe while – we have just one minute. We could look at – because I know you provided some information about missing data. That's in 2B 7.2. So, I appreciate those. And actually, I just also want to thank you for putting the information, the relevant information with the questions. It helps go through this, so I appreciate it.

OK. If you want to just make any overall comment, we don't have to go through these in detail, but just any overall comments we feel about missing data and whether that's an issue for any of the domains.

Lucille Schacht: We see very little missing data. The place that usually jumps out is the questions related to the participation and treatment in relation to working with their next care provider. So, if people are completing this during the annual review, those questions aren't relevant and so that has a lower response rate on that particular item. And then also, there are some two medication questions and not everyone is on medication, so that may have a lower response rate as well.

And basically, that's what we see is the missing data tends to be related to the discharge and interaction with the next care provider and it's equivalent to the

– around the level of people who are actually doing these surveys on annual review.

Karen Pace: OK.

Sharon Cross: This is Sharon. Can I just interject a question here for a second?

Lucille Schacht: Absolutely.

Sharon Cross: When you were – I'm curious and maybe buried somewhere in here that I missed the data. Is there anyway to know what the percentage or what the rates are of how many people are completing this at the annual review versus when being discharged from the hospital? And what are organizations defining as an annual review?

Because we've had some comments in the past and I know the comment that was sent in regarding annual reviews in my mind usually happen as an outpatient service versus an inpatient? So I'm kind of curious about the connection between those things.

Lucille Schacht: Sure. So an annual review – many of our facilities run long-term care units. And so in order to really get the consumer voice, they do need to do a survey other than discharge and so annual review is really helpful for that.

We get around 25,000 surveys a year. But only a small percentage of those are actually annual review. I don't have the number off-hand. We have looked at differences between the annual review responses and discharge responses, that's something that we do provide for additional information back to a facility using this information. So, we do look at it and it allows for those facilities that have long-term care units to get consumer feedback.

Sharon Cross: And that's helpful to know. It helps in my mind just to know that it's a lot less as far as the annual review. Is there a reasoning or thought behind, because you mentioned the possibility or discussions about an outpatient survey? Is there reasoning behind why the same – again, I'm just thinking about what the experiences are that patients have that are inpatient for behavioral mental

health issues. Is there reasoning behind why it would be this survey versus an outpatient specific survey at the annual review?

Lucille Schacht: Because they're still an inpatient. And so it's about how their inpatient experience is going about the interactions with their staff and how they're feeling in terms of their treatments, their interactions, whether they feel like they're making progress since their last review.

Sharon Cross: So you're actually talking – wait, so you're talking about inpatients, people that are in long-term care for more than a year timeframe?

Lucille Schacht: People in the hospital – psychiatric hospital maybe in for more than a year, yes.

Sharon Cross: But then if they're in for more than a year, they really wouldn't be – and so they are evaluated – OK. So maybe I missed this. So, why would they be given one to begin with at a discharge then – and then they'd get it again in a year if they're not being discharged?

Lucille Schacht: They would get it at their annual review. If they're in for more than a year, they would get it at their annual review as a way for the facility to stay informed about the patient experience of their care.

Sharon Cross: OK. So it's not actually tied specifically to a discharge for those that are long-term patients?

Lucille Schacht: Well, for a long-term person, they would also fill one out at the time of discharge.

Sharon Cross: OK.

Lucille Schacht: If their annual review and their discharge are a week apart, they probably won't get the survey twice and that would be a decision locally at a hospital.

Sharon Cross: Got it.

Lucille Schacht: But if it's six months apart then they would probably give them the opportunity to do the survey again.

Sharon Cross: Got it. OK. That helps. Thank you very much for the clarification with that. I appreciate it.

Lucille Schacht: You're welcome.

Lauralei Dorian: OK. Are there any other questions about this measure? I'm just looking at the time and I want to make sure the last measure, we get to that.

All right. Thank you, Lucille.

Lucille Schacht: You're welcome.

Lauralei Dorian: Now, we are moving to 006 which is the consumer – the CAHPS Health Plan Survey Version 5.0 Medicaid and Commercial and this is a standardized survey instrument which asks enrollees to report on their experiences access in care and health plan information and the quality of care received by the physician. I believe it's made up of eight separate measures.

So, we'll go to the evidence section. And again, we'll ask whether the relationship between the measured PRO and at least one healthcare action is identified and supported by the stated rationale. We'll just see what …

Karen Pace: Right.

Lauralei Dorian: … the developer with that.

Karen Pace: Right. And again, we had asked for this to be done for the measures. And I see that in this diagram, those are in the, I guess middle box. Well, we know that the measure titles are – no.

So is this – we'll ask the developer to explain the relationship of this to the specific eight measures that they're submitting for endorsement.

Chris Crofton or someone else from the CAHPS group?

Anne Monroe: Actually …

Female: Anne, do you want to address this or do you want the Westat group to talk this through?

Anne Monroe: Why don't the Westat people handle it?

Female: OK.

Female: OK.

Karen Pace: And I see in the red boxes where you have your measures, right, or some of them?

Female: And what we were trying to do with this graphic was really show how, you know, the structural features of health plan such as the type of health plan and the market conditions, you know, lead to some variability in the measure.

So, in that red box, do you see the four composites listed getting Medicare, getting care quickly, how well doctors communicate, and health plan information and customer service. And then in addition to those composite measures, we have four global or overall readings of the healthcare received that patients' personal doctor, special listing and the health plan. So that encompasses the eight measures there.

And we show in this graphic and in a literature review that follows how these measures are related to other measures of clinical quality, process measures, treatment measures, procedures measures, and also healthcare related patient behavior. There has been especially strong (inaudible) on adherence to medication. But we also discussed adherence to care and visits, patient activation, self management, use of preventive care and inpatient engagement in the decision making process.

And then finally, through that link, we get to better patient outcomes. So, better patient reported health functioning well-being, better clinical outcomes, rest unnecessary utilization, lower cost and therefore resources and lower rates of complication or adverse events. So we're really showing you kind of the lifespan of the things that cost CAHPS measures and how they're related to other outcomes.

And I know that in prior committee meetings, you have been interested in interventions that could be used to improve CAHPS scores. We didn't discuss that in this particular document but we do discussed it in our main measure submission form in more detail about how we know that CAHPS had been used as part of quality improvement.

Karen Pace: OK. And just to clarify in terms of our – the way our current criteria are setup, what we were most interested in here for this and this is great to show that these things are all related but the things that are in the clinical quality box, the process treatment and procedures that really can impact the patient's experience but we'll, I guess, be seeing that somewhere else in the submission form.

OK. Any comments or questions from the committee?

Lauralei Dorian: OK. We can move on to 1B which is performance gap. Let's see if we can bring that section up. So, just the performance data provided demonstrated gap of care. All right, there are demonstrated disparity – other demonstrated disparities for certain populations or group.

Karen Pace: So, in 1B2, we asked the developer to provide the distribution of scores for the entities being measured and also to provide any information about over time and they're basically referred as to another place. So, I'm not sure. The top box scores indicate room for continued improvement, the highest top box score with 74 percent. So, I'll just stop there and see if any of the committee was able to chase down that information.

Female: Right. And just to point out for the committee and the NQF staff, because it's such a large volume of data and we were asked not to put too much in these forms. We did put our current performance data in an Excel spreadsheet which you should have.

And we also report the measures overtime as part of another question on improvements later on. So that information is maybe note quite as accessible as you would like but …

(Samuel Burner): Right.

Female: … we can help you find it if that would be helpful.

Karen Pace: Yes, we'll have to, I guess, search for that. We try to get at least a good summary information in here for the eight measures but – because – but let's go – we'll move on. Same way then, I guess, with the disparities information. Do you have any information?

Female: We do and that disparities information is in the Excel spreadsheet.

Karen Pace: OK.

Female: Now, that is available.

Karen Pace: OK. Let's move on. Any questions from the committee about this? I don't know if we'll have to try to help you find that. I don't think we have time to search through that Excel file while we're on the call here. So, why don't we move on to 1C Lorelei which is high priority.

Anne Monroe: Could I just ask one question? This is Anne.

Female: Yes. Sure.

Anne Monroe: Have you seen, you know, if the healthcare delivery system gets reformed and the plan's relationship with the provider system changes, do you – have you seen or looked at any amendments to health plan CAHPS that you think are more appropriate given the certain – given the today's delivery system and the relationship between the two?

Female: We're aware that CMS is developing new versions of CAHPS that may help to address this issue in particular for ACOs which are supported by the Affordable Care Act. That's all we know right now. And our database doesn't really stretch far enough into the future to look at any, you know, to really get a hand on in the impact of healthcare reform on the scores as the measures currently specified.

Chris Crofton: Thank you. This is Chris Crofton. In addition to that, I'd say that the team is considering some changes to the measures to make them fit better with changes in healthcare delivery. Some possible changes might be in the communications area, also in the area of access since there are many different ways now to contact your doctor or your doctor's office or the health plan then were available when these surveys were originally developed.

So I can't say anything definite about that right now. But we're getting feedback from people and our plan is to respond to it as best we can.

Female: Thank you. I would think care management might look different overtime as well if a lot of that gets validated to the health – to the provider.

Chris Crofton: Yes.

Female: And it used to be more in the health plan purview. But I don't want to take more time on this. Thank you for clarifying.

Lauralei Dorian: Right. And now, we're looking at 1C5 which is evidence that the target population values the measured PRO and finds it meaningful.

So, it looks like the developers provided information maybe if you just like to walk through on how you involve patients in the development of the measure?

Chris Crofton: I'll turn to Westat for that explanation since they were critical in developing this package.

Female: OK. Thanks, Chris. CAHPS health plan is a relatively old survey and we've listed some of the steps that have contributed to – about the original design …

Female: Could you speak up, please?

Female: Yes. Can you hear me?

Female: Much better.

Female: OK. And so we've listed here some steps which contributed to the original development of the survey and also its refinement overtime. We have done

focus groups with consumers and, you know, developed our measures in consultation with technical experts, we've done cognitive testing, field testing, had a public comment period. And we can also say that at this point in CAHPS lifespan, there is a published literature out there that indicates that patients are using these measures.

So, I think it's the second paragraph, you'll see under the bulleted list that discusses how, you know, people are using CAHPS in selection of managed care plans.

Lauralei Dorian: OK, great. And now, we will move on to 2A1 and 2B1 which is specification. Let's bring that up.

Karen Pace: So we'll go to the – yes S4. It's where the numerator statement and that – because there's some general comments. So at six numerator details, it's where we had asked for the specific measures and the items that go into the measures. So, I don't see those there though the calculation steps are there. Did you put those in the data Excel file, the data dictionary? The CAHPS team? The …

Female: Yes, ma'am. You're talking about the list of each individual item within each composite?

Karen Pace: Yes.

Female: Yes. That's in the Excel file.

Karen Pace: OK. All right.

Female: I apologize for it not being more readily available but we were going by instructions that said we couldn't put anything longer than a page.

Karen Pace: That's fine. That's fine.

Yes. For the – for any kind of coding, we – if it's going to be longer than a couple of pages, we do suggest those going in the Excel files, so that's good.

And it looks like you're doing the same top box scoring. Is there anything different about the scoring that you do in this than some of the other CAHPS?

Female: No. This is very standard.

Karen Pace: OK. So, we'll move onto testing now.

Lauralei Dorian: So, 2A2 is reliability testing and to 2B2 is validity testing. And again, we want testing to be – and have been conducted at both levels with an adequate scope to generalize for widespread implementation and within appropriate method. So, we'll bring up the reliability first.

And so if the developer could walk us through your testing approaches and how the different testing approaches for the two different level scale and facility levels for these.

Female: So like the other developers, we are showing Cronbach's alpha for the individual level showing how the internal consistency reliability of the measure and showing how well the items hang together.

And for our site level reliability which we're calling plans reliability for this particular survey. We've provided the formula for you. I believe it's out on the next page. There you go. Yes. And so this formula is allowing us to look at between and within groups of variability and showing how well they plan to hang together.

Karen Pace: So, and then we can go down to the results in 2A2.3 and we want to – just like any general observations about these.

Female: So we did note in our write up that, you know, the typical thresholds that is used to judge these is .7 and as you can see one of our measures is pretty significantly below that. But we would note that this measure was designed primarily with health plans in mind. And so we felt that the plan, the site reliability, plan level reliability was going to be our better measure of the reliability of these items.

Karen Pace: So, the next table talks about health plan child Medicaid version. So do we actually have more than eight measures? Is there child versions of – how many of these? Is it …

Female: There's two versions of this survey. There's the adult version and the child version. But the composites are the same. There is one additional item in one of the composites which is specified in the Excel spreadsheet. But in general, they were looking at the same composite, the same measures to different populations.

Karen Pace: OK. Thank you. All right. And it looks like then for the child, these are a little – are less than for your adult version. Do you have any thoughts about that?

Female: I believe our mean respondents for plan was a little less down in the child section for a couple of those items. Oh, that's not true. So that's not – let's see. So you're saying that the …

Karen Pace: I was just looking at the Cronbach's alpha.

Female: Individual, not the …

Karen Pace: They're just – yes. They're just a little bit lower not significantly. I was just looking back and forth. I see.

Female: Yes. I don't really have a comment on why that …

Karen Pace: OK.

Female: … might – OK. Sorry there. I got confused between the …

Karen Pace: No. That's OK. Let's go on down to the plan reliability then. All right. And I guess the one that – let's scroll down, the one that stands out is this global reading of specialists.

Female: Right. And here is where my comment now does make sense about …

Karen Pace: OK.

Female: … the lower number of respondents per plan. So individuals who haven't seen a specialist in the past year don't answer those questions and you have a lower sample size per plan for those items. We think that's contributing to the lower value.

Karen Pace: So do you have a – any – has it always been that low, or what's the plan for that then?

Female: Yes. It has typically been lower. At present, we don't have any specific plans to modify the items.

Karen Pace: Comments or questions from the committee members? Any …

(Leigh Partridge): This is (Leigh). I singled out the specialist issue throughout this submission and was curious. Is that – has that question been answered from the very beginning?

Female: Yes. Yes.

(Leigh Partridge): It has. And has it always been low like this and have you thought of dropping it?

Ron Hays: This is Ron Hays. It's always been in – and we've always seen that it applies to a lower number of people which make sense. But …

(Leigh Partridge): Right.

Ron Hays: … generally, we do our overall sample size recommendations based on knowing that. So in the beginning, we recommend 300 completes for a plan with the idea that some items would only apply to a third of the people. And so, as long – the key is to have enough reliability for the worst case items. So as long as you have a sample size that will give you a reliability of .70 at the health plan level, it's OK. It's just a function of applicability of the item. So you have to have a large enough sample to make sure you get enough people to answer it.

(Leigh Partridge): So then is this used in anyway? Because it says that the plan level of reliability.

Ron Hays: Well, this is from, you know, the data that was collected here for actual applications of CAHPS. We, you know, make sure people know that they have to estimate with the reliability. It is based on the sample size they get and we provide recommendations for how they can get enough based on what we know. So that's why – the general rule was, you know, at least 300 completes for a plan, so you'll have 100 completes for any item. And then, you know, that map to enough reliability of .70 level reliability at the plan level.

So we have all the information that goes into the guide book. It says, "OK. If you want enough reliability to report this, you have to have this starting sample, because we know you're going to lose cases." So these data are used as a basis of providing the recommendations for what the sample needs to be when these are reported.

(Leigh Partridge): Right.

Ron Hays: But we, you know, we can have a sample that isn't sufficient for reporting and still estimate what you need in the real application when you want to report that item.

(Leigh Partridge): I guess my question was just on the reliability. It just seemed to me that this particular measure on the – back in the rationale for its importance and the evidence behind it. And also the fact that a lot of people, they haven't seen a specialist so it does lower your end. I wondered whether you had considered along the way deciding this was not really yielding data that was all that useful.

Ron Hays: I don't think that's ever been the case. I mean, I think …

(Leigh Partridge): OK.

Ron Hays: … people think specialty care is really important and some might, you know, a minority of people that it's still a substantial number and when they need it they want to know what people think of it.

Karen Pace: All right. OK. So why don't we go into validity.

Female: OK.

Female: Move on to 2B. The validity testing. So, we're looking at – again, whether the testing was conducted at both levels with adequate skills and method. So, we'll ask the developer again to explain this approach for a validity testing, please.

Female: Ron, did you want to speak to that or you want the Westat?

Ron Hays: No. I think you're doing a fabulous job.

Female: C'mon.

Female: That's a good endorsement.

Female: Yes.

Female: So, we did a couple of things. We looked at the relationship between each composite and the global measures of patient experience that (inaudible) with them. So, as one example, just like why would be think this is in accordance? Either composite measuring how well respondents, personal doctor communicates, we expect to be pretty strongly related to their overall reading of their personal doctors. So, you're finding that kind of relationship supports the validity of the measure.

So in table 2B2.3 in that table showing relationships between each of the composites and the global …

Female: Could you speak up a little bit? We're – you're kind of fading out.

Female: Sorry about that. Just pointing out that these results are in table 2B2.3 which now you've got up there, showing you relationships between each of the

composites in the global rating. And then we also examined Spearman rank order correlations among the composites. You know, we expect that different dimensions of patient experience are related to one another but we're watching out for cases where they might be too highly correlated, because in that case they would be measuring the exact same thing and wouldn't really be considered separate measures.

So, in evaluating the figures that you see here, you know, we were satisfied that we had significant correlations between our composites in the global rating. And we were in general, you know, satisfied with the strength in – of the relationships among the different composites.

Karen Pace: OK. And then at the scale level, did you do any validity testing or – is this actually – this is at the plan level, right? The computed score levels, the 2B2.3?

Female: That's right. This is at the plan level.

Karen Pace: OK.

Female: Scale to scale and scale to global.

Karen Pace: And did you do at the patient level validity testing such as this?

Female: We did not.

Karen Pace: OK. All right.

Lauralei Dorian: So are there any questions from the committee or comments about testing?

I'm looking at the time. Are there any questions about this process in general, or feedback, or reflections you'd like to share in the call?

OK. So before we end, we'd just like to open the line for public comment. I believe, (Amy), the line is open to everyone right now, right?

Operator: Yes, ma'am. All lines are open.

Lauralei Dorian: Great. If there are any members of the public who'd like to make a comment, please do so now.

All right. Well, thank you everybody. Thanks to the committee members for dialing in and participating today and to the developers for being able to respond to questions and explain your measures.

Our next steps will be to summarize these workgroup calls and send them to you. And as you know, we have our in-person meeting here on July 28th and 29th in Washington D.C. So you should have received travel information from NQF travel department, but if you have not yet receive that, please let me know so that we can make sure that you can book your accommodation and your flight as quickly as possible.

And I'll just ask one last time if there are any questions about the next step?

All right. Well, thank you again and have a great …

Karen Pace: Thank you everybody.

Female: … afternoon and good Fourth of July. Bye.

Female: Do you know – you want to link her, or?

Female: Yes. Yes. Let's link her.

Female: OK.

Female: I'm not sure if she'd be able to see us.

(Off-mike)

Female: Yes.

END