NATIONAL QUALITY FORUM

+ + + + +

PERSON AND FAMILY CENTERED CARE PHASE 3

+ + + + +

MONDAY

JUNE 6, 2016

+ + + + +

The Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Lee Partridge and Christopher Stille, Co-Chairs, presiding.

PRESENT:

LEE PARTRIDGE, Co-Chair
CHRISTOPHER STILLE, MD, MPH, FAAP, Co-Chair
BETH AVERBECK, MD, HealthPartners, Inc.
KATHERINE BEVANS, PhD, Children's Hospital of Philadelphia
SAMUEL BIERNER, MD, University of Nebraska Medical Center
REBECCA BRADLEY, LCSW, HealthSouth Corporation
JENNIFER BRIGHT, MPA, Momentum Health Strategies
DAVID CELLA, Northwestern University Feinberg School of Medicine*
SHARON CROSS, LISW, Ohio State University Wexner Medical Center
DAWN DOWDING, PhD, RN, Visiting Nurse Service of New York*
NICOLE FRIEDMAN, Kaiser Permanente
STEPHEN HOY, Patient Family Centered Care Partners
SHERRIE KAPLAN, PhD, MPH, University of California Irvine School of Medicine

LINDA MELILLO, MS, MA, CPHRM, CPXP, Spaulding Rehabilitation Network, Partners Healthcare System
LISA MORRISE, MA, Patient & Family Engagement Affinity Group, National Partnership for Patients
ELIZABETH MORT, MD, MPH, Massachusetts General Hospital/Massachusetts General Physicians Organization*
LENARD PARISI, RN, MA, CPHQ, FNAHQ, Metropolitan Jewish Health System
DEBRA SALIBA, MD, MPH, UCLA/JH Borun Center for Gerontological Research, VA GRECC, RAND Health*
LISA GALE SUTER, MD, Yale School of Medicine, Yale Center for Outcomes Research & Evaluation (CORE)
PETER THOMAS, JD, Powers, Pyles, Sutter & Verville, P.C.
CARIN VAN ZYL, MD, FACEP, Keck School of Medicine of University of Southern California

NQF STAFF:

HELEN BURSTIN, MD, MPH, Chief Scientific Officer

ANN HAMMERSMITH, JD, General Counsel

ELISA MUNTHALI, MPH, Vice President, Quality Measurement

ALEXANDRA OGUNGBEMI, Project Analyst

DESMIRRA QUINNONEZ, Project Analyst

KIRSTEN REED, Project Manager

SARAH SAMPSEL, MPH, Senior Director

SUZANNE THEBERGE, MPH, Senior Project Manager

ALSO PRESENT:

SVEN BERG, MD, MPH, CPE, West Virginia Medical Institute/Quality Insights
NICHOLAS CASTLE, PhD, University of Pittsburgh Graduate School of Public Health*
DEXANNE CLOHAN, MD, Retained Consultant
KATHLEEN DANN, UDSMR
MAGGIE DIVITA, UDSMR*
DAVID GIFFORD, MD, MPH, American Health Care Association
DANIEL GREEN, MD, FACOG, Centers for Medicare & Medicaid Services*
DEANNA HAYES, PT, DPT, MS, Focus on Therapeutic Outcomes*
BETH JACKSON, PhD, Truven Health Analytics
KERRY LIDA, PhD, Centers for Medicare & Medicaid Services
JANE LUCAS, Quality Insights of Pennsylvania
CORETTA MALLERY, PhD, American Institutes for Research
JAMES MULLER, American Health Care Association
PAULETTE NIEWCZYK, PhD, MPH, UDSMR
GARY REZEK, West Virginia Medical Institute*
ANITA SOMPLASKY, RN, West Virginia Medical Institute/Quality Insights of Pennsylvania

*** present by teleconference**

A G E N D A

P-R-O-C-E-E-D-I-N-G-S

8:39 a.m.

CO-CHAIR PARTRIDGE: Welcome again to Washington, D.C., on a typical June day, temperature nearly 90.

So just because I suspect you, like me, wondered what all the construction is across the way, that was for many, many years, as long as I lived here, the site of the Washington Post. But its new owners decided to sell it, and the property is now being developed, I understand, into a mega-complex of residential and commercial.

MS. HAMMERSMITH: Fannie Mae.

CO-CHAIR PARTRIDGE: Oh, Fannie Mae is the principal tenant? Okay, so the new home of Fannie Mae. It seems very strange because for years we've been looking at those sort of yellowish brick buildings and watching Post reporting.

It's Chris and my pleasure to welcome you here today, and Sarah is going to start us

off with some of the ground rules, et cetera. Right?

MS. SAMPSEL: Yes. So, good morning, and thank you all. I know a number of us had problems making it across the country yesterday, and we'll have a couple of people joining us a little bit late, but we do have a couple of folks on the phone, as well, so just want to remind folks that we will have some Committee members on the phone.

But I'm Sarah Sampsel, I'm the Senior Director for this project. And how we're going to start this morning is by doing introductions and disclosures of interest. So we'll have staff do introductions first, and then we'll turn things over to our General Counsel, Ann Hammersmith, who will go around and give you the instructions for disclosures of interest.

But again, just want to welcome you all and do hope that you will get a lot out of this meeting. We are very much looking forward to all of your feedback, and we'll talk about

some of the logistics and other things after we do our introductions. So why don't we go ahead and start. Desi, do you want to go ahead and start, please?

MS. QUINNONEZ: Good morning, everyone. I'm over in the corner. I'm Desmirra Quinnonez. You'll see me often as Desi on my emails, and I'm the Project Analyst on this project.

MS. REED: Good morning, everyone. I am Kirsten Reed. I am a Project Manager here at NQF. Today actually marks week seven for me, so I am very excited to really jump right in and get started with everyone today.

MS. THEBERGE: Good morning, everyone. I'm Suzanne Theberge. I'm the Senior Project Manager on the team, and enjoying working with you all again.

MS. OGUNGBEMI: Good morning. My name is Alexandra Ogungbemi, and I am a Project Analyst at NQF. I just work on the quality measurement department.

MS. MUNTHALI: Good morning. I'm Elisa Munthali. I'm Vice President for Quality Measurement; just wanted to welcome everyone and thank you for serving on the Committee.

MS. HAMMERSMITH: Hi, everyone. I'm Ann Hammersmith. I'm NQF's General Counsel. Many of you have heard the little talk I'm going to do before, and I'll do it again just to remind everyone.

If you recall -- I'm sure you do, it's a rather long form -- we sent you a conflict of interest disclosure form a while ago in which we asked you some detailed information about your professional activities. So what we're going to do this morning is go through an oral disclosure, the purpose of which is transparency so that each of you and the public who are present or listening know where you're coming from.

I'll remind you that you sit on the Committee as an individual. You're here because you're an expert in this area. You don't represent your employer, you don't represent

anyone who may have nominated you to serve on this Committee.

Some guidelines for what you should disclose. You should only disclose things that are relevant to your service on Committee. In other words, that are relevant to the subject matter of the Committee. We are particularly interested in speaking engagements, consulting, grants, research. But again, only if it's relevant to the subject matter that's before the Committee.

Just a reminder that just because you disclose doesn't mean that you have a conflict of interest. As I mentioned before, the idea behind this is transparency. So you might be involved in something that's relevant to the subject matter but it's not a conflict of interest, but you may wish to disclose it because it is relevant to the subject matter.

So with that, I'll start with the Co-Chairs. We'll go around the table. If you're on the phone, I will call on you at the end of the

in-person disclosures.

CO-CHAIR PARTRIDGE: I'm Lee Partridge. I retired on the 1st of April from many years of volunteer service with the National Partnership for Women and Families, here in Washington. And I picked up a new part-time assignment in my new home, which is now New York City, with the United Hospital Fund, a 125-year old nonprofit foundation in the city where I am working happily on pediatric care, which is my first love. And I have no conflicts or other issues to disclose.

I do want to say, however, as we go around the room, particularly since we have seven new members, not all of whom are here, but most of whom are here, just if you would add a sentence or two about where you're from just so that we have a better -- just a little bit better acquaintance.

CO-CHAIR STILLE: Good morning, everyone. My name is Chris Stille. I'm a professor and section head at General Academic Pediatrics at the University of Colorado School

of Medicine and Children's Hospital Colorado. It's an honor and a pleasure to be working with all of you. My academic interests are in improving systems of care for children and families with special care needs. I have no disclosures or conflicts to note today.

MEMBER THOMAS: My name is Peter Thomas. I think this is the third time I've been here. I'm happy to be here. I'm a consumer. I have two artificial legs that I kind of wear on my sleeve because it's given me a tremendous front row seat to the healthcare system throughout my life, and kind of went into that field, I'm a lawyer and a lobbyist and do a lot of healthcare work from a disability and rehabilitation perspective.

So I represent a lot of different clients in the rehabilitation space. I don't represent UDSMR, which is one of the measure developers. I don't believe I have any conflicts to disclose. I will be speaking to the UDSMR in August, but it's not really about -- it may touch

on quality issues but it's really about rehabilitation issues overall. And I've never worked with anyone on measure development, so I don't think I've got any conflicts.

My office is two blocks away and I'm from Washington, D.C., and Chevy Chase, Maryland.

MEMBER MELILLO: Hi, I'm Linda Melillo. I am the director of quality and compliance at Spaulding Rehabilitation Hospital Cape Cod, and I'm also the Director of Patient Experience for the Spaulding Network, which is part of the Partners Healthcare System out of Boston.

I suppose I don't -- I don't know if you'd call this a conflict or not, but Spaulding Rehabilitation Network includes inpatient rehabilitation facilities, long-term acute care, skilled nursing facilities, and home care. So I just wanted to get that out there.

I'm thrilled to be here. I'm also a consumer. I had a traumatic brain injury years ago, so lots of use of the healthcare system. So great to be here, thank you.

MEMBER BRIGHT: Hi, everyone. I'm Jennifer Bright. I have my own consulting company, Momentum Health Strategies, and I'm based here in Alexandria, Virginia. I'd like to think I have a place at this table today because I grew up on patient advocacy working in the mental health community and started a coalition about 12 years ago, back when it was not fashionable to talk about patient engagement.

So I started to coalesce a whole bunch of patient groups, from kidney to diabetes, to get involved and to educate their members about why patients' rights was important and comparative effectiveness and policy development. And currently, my clients are mostly non-profits, but I also some advisory to -- not currently, but in the past with some healthcare companies on topics related to policy development, patient engagement, which is my passion, as well as care integration for behavioral health and primary care. So thanks.

MEMBER MORRISE: I'm Lisa Morrise, and

I'm from Salt Lake City, Utah. I am a patient advocate. My daughter, Kirsten, who's now 23 and a student at Utah State studying social work, was born unable to breathe or swallow. And her medically intense journey that included a tracheotomy, NG tube, and 45 surgeries to-date has helped me understand the needs that we have for improvement in patient engagement and patient and family centered care in the medical system.

And while we feel like we've been extremely fortunate with the providers who have worked with Kirsten, especially as I work with other parents whose children have the same birth defect but are not getting the kind of care that they need. We found that in general the system really needs patient input. So now I work as a subcontractor to a subcontractor for CMS and work in patient and family engagement on projects like Partnership for Patients and TCPI, and also help hospitals around the country develop patient and family advisory councils, and just in general consult in other areas of patient and family

engagement. And I don't believe I have anything to disclose. Thanks so much.

MEMBER BEVANS: Good morning, everyone. I'm Katherine Bevans, I'm an associate professor at the University of Pennsylvania School of Medicine and the Children's Hospital of Philadelphia. I'm a researcher who develops patient-reported outcome measures, especially for use with children and parents. And I also do work on the development of methods for enhancing patient, family, other stakeholder engagement in research. I don't believe I have any conflicts of interest.

MEMBER VAN ZYL: Hi, everyone. I'm Carin Van Zyl. I am Assistant Professor of Medicine at the Keck School of Medicine. And I run the programs for palliative care at LA County Hospital and the Keck School. I have no conflicts to disclose.

MEMBER HOY: I'm Stephen Hoy. I'm a new member here. I work in a place called Mammoth Lakes, California. I'm the director of strategy

and programs for a small consulting company called PFCC Partners that partner with healthcare organizations to advance PFCC practices in their organizations. And I don't think I have any disclosures, conflicts of interest to disclose. Thanks.

MEMBER CROSS: Good morning. My name is Sharon Cross. I am currently the Program Director for patients care at the Ohio State University Wexner Medical Center. And within my role I am the person that manages all of our patient and family advisors and advisory councils, and work with researchers who are engaging patients and families in their research design. I have participated in some different projects with PCORI but I don't believe I have any conflicts of interest to disclose. Thank you.

MEMBER AVERBECK: Good morning. I'm Beth Averbeck from Minneapolis-St. Paul, Minnesota with Health Partners. We're an integrated care delivery and financing organizations. My administrative

responsibilities, I lead primary care, which is about 400 physicians. And my clinical practice is in long-term care, TCU, assisted living. And I've also worked with Minnesota Community Measurement on some patient-reported outcomes and electronic health record measure submissions. Thank you. Oh, I have nothing to declare.

MEMBER PARISI: Good morning. My name is Len Parisi. I am out of New York City, vice president of quality management for MJHS. I'm a post-acute care provider, home care hospice, skilled nursing facility. I have no conflicts of interest but I do sit on the Joint Commission Standards and Survey Methods Committee, as well as the Home Care Advisory Committee.

MEMBER BRADLEY: Good morning. My name is Becky Bradley. I'm a social worker by training. I'm currently the National Director for Case Management for HealthSouth Corporation which is a post-acute provider. We are users of many instruments in our settings, and I have nothing to disclose.

MEMBER KAPLAN: I'm Sherrie Kaplan. I'm a Professor of Medicine and Assistant Vice Chancellor for Healthcare Measurement and Valuation at University of California Irvine School of Medicine. I have currently a grant from PCORI to further the development of the children's measure of self-reported functional status. It's animated touchscreen-based. The functional status measurement was developed about 1996 and is being developed now for anesthesia and perioperative anxiety, post-operative pain management.

And I also developed a number of measures in patient-reported outcomes, and participatory decisionmaking style that's relevant. Those two things are relevant but they're not -- for me don't constitute a conflict of interest.

MEMBER BIERNER: Good morning. I'm Sam Bierner, Professor and Chair of the Department of Physical Medicine and Rehabilitation, University of Nebraska Medical Center. And I'm medical

director of the Madonna Rehabilitation Hospital in Nebraska. And I have no conflicts of interest.

MEMBER FRIEDMAN: My name is Nicole Friedman, and I am the medical anthropologist in my background, and currently I am the Regional Manager for Patient Navigation and Community Health Work for Kaiser Permanente in the Northwest. I am from Portland, Oregon. I'm new, and I have no conflicts of interest.

MS. HAMMERSMITH: I'm now going to call on the people who are on the phone. Dawn Dowding. Is Dawn Dowding on the phone?

MEMBER DOWDING: Yes, hi. Sorry, I was on mute. I'm Dawn Dowding. I'm a professor at the School of Nursing at Columbia University, and I have a joint appointment with the Visiting Nurse Service of New York, and I have no conflicts.

MS. HAMMERSMITH: Thank you. David Cella.

MEMBER CELLA: Good morning. This is David Cella. I'm Professor of Medical Social Sciences at Northwestern Medical School. This is

my second time on the Committee. I don't have any conflicts that are directly related to any of the submissions for this meeting, but I do have a grant from NIH to curate and distribute PROMIS and Toolbox and Neuro-Qol measures for use around the country. And some of those at some point may be used as performance measures.

And I'm co-investigator on a PCORI grant to the Rehabilitation Institute of Chicago to develop performance measures for post-acute -- patient-oriented performance measures for post-acute care, none of which are in front of the Committee at this meeting.

MS. HAMMERSMITH: Okay, thank you. Debra Saliba.

MEMBER SALIBA: Hi, I'm Deb Saliba. I am a geriatrician and health services researcher, Professor of Medicine at UCLA, the VA, and the RAND Corporation. I don't have any conflicts of interest. I am, as a researcher, funded by multiple federal agencies and federal foundations, but none of my current work is

developing quality metrics.

MS. HAMMERSMITH: Okay, thank you.

MEMBER SALIBA: And I look forward -- I apologize that I can't be there today and look forward to participating.

MS. HAMMERSMITH: Okay, thank you. Liz Mort.

MEMBER MORT: Good morning. This is Liz Mort from Mass General. I'm the Senior Vice President of Quality and Safety here. I have a background in health services research and quality measurement.

My apologies that I'm not there in person and that I can only participate partially today. My clinical and administrative responsibilities changed dramatically and unexpectedly.

I have no conflict of interest other than the fact that I am from Mass General, and tomorrow you will be considering a measure from MGH. Should there be decisions that need to be made post the in-person meeting, I will recuse

myself from those decisions. And happy to be participating again.

MS. HAMMERSMITH: Okay, thank you. Is anyone else on the phone who I've missed?

All right. Thank you, everyone, for those disclosures. You have very interesting backgrounds. This is one of the most interesting committees I've done disclosures for.

Anyway, do you have any questions of each other or anything you want to discuss, or any questions for me?

Okay, before I leave you, I just want to remind you of one thing, which is if at any point in the meeting you think you may have a conflict, or if you think someone else has a conflict that they haven't declared, or if you think someone is going beyond stating their opinion and heading into bias, we ask you to speak up.

What we don't want is to have you pop up three months later and go, "you know, I think I may have had a conflict" and you didn't tell us

at the time of the meeting. So if any of the things I've described occur, in your opinion, you're always free to bring it up in realtime in the meeting. If you don't want to do that, you could approach your Chairs who will work with NQF staff, or you can approach staff directly. So with that, I will let you get on with your meeting. Thank you.

MS. SAMPSEL: Thanks, Ann. And thanks, everyone. A couple of housekeeping issues. First of all, we do have breaks put into the agenda. However, if at any point you still feel you need to stand up, do whatever, the restrooms are out through the lobby and to the right, but we do have breaks at 10:45, and then lunch, and the one in the afternoon, as well.

You should have all hopefully been able to log into our guest network, but if not, the information is on the screen right now. We ask you to mute your cell phone during the meeting. We definitely understand if folks need to get up, but if you have to have a

conversation, if you could leave this area. We do tape these conversations in these meetings, and then we also have a transcriber in the back corner of the room who is documenting everything. So, microphones.

MS. THEBERGE: Yes, there's somebody on the phone, I think, who also has the webinar streaming. If you could mute your computer if you're dialed into the open phone line, we would appreciate it. Thank you.

MS. SAMPSEL: Suzanne can fix echos; I can't. So with that -- and, again, somebody on line does have their computer streaming audio. The microphones, when they're lit it means that they're active. Only two can be active at one time, so we ask that as you speak, once you're done if you'll turn your microphone off, and all of us need reminders about that typically.

And then the way we do it, if somebody wants to speak, if you have a question or a point to make, or whatever, if you will raise your card like this, that way Chris and Lee can identify

who wants to speak next.

So with that, we want to do just a very brief overview, again, of what we're trying to accomplish today and do some reminders. But also recognize the fact that we've had a couple of calls already. I think we've had a lot of discussions about the criteria. But just a reminder, in this meeting we have one maintenance measure for review and then we have 12 -- some with multiple submissions on one form -- but 12 additional new measures.

So the bulk of our work are new measures. But as a reminder on the maintenance process, these are all measures that are currently endorsed, so they've already been evaluated to meet the criteria in the past. However, you know, NQF criteria have changed over time, so we ask you to look at the measures in light of the criteria that have changed, and the fact that you all have been actually brought up in this process with the new criteria, so you may have additional questions.

For some of the criteria, such as evidence and testing, if there is no new evidence or data, you're able to accept the past evaluation without further discussion and voting. However, the measure that we're looking at today, there's been some updates to specifications, et cetera, and therefore, we are going to have those conversations about evidence and testing.

And we want greater focus placed on the current performance and opportunity for improvement. Obviously, you know, one of the goals of measurement is to improve over time and to really show that it's making a difference. And so that's why we want some focus on the performance and if there continues to be growth in the measure. And then usability and use, how well is the measure being used, who is it being used by, et cetera. Next slide.

These are just reminders. And, again, we went through these on the conference call. The SDS, or the sociodemographic trial, we convened an expert panel on demographic factors. And prior

to this panel, NQF was not asking folks -- and, in fact, suggested against including SDS factors in performance measures.

There are two perspectives on SDS. One is that when you adjust it will mask disparities. And the other is that you have to adjust because it's necessary to avoid making incorrect inferences about the context.

So this expert panel came together and said, yes, we acknowledge that, but we still feel there is a need to understand more about what is going on. You know, for each individual measure perhaps there needs to be more consideration on how the SDS factors are impacting that measure.

So the SDS panel recommended, the NQF Board approved: a two-year trial period with adjustment of measures for SDS factors is now no longer prohibited.

So, basically, this started in April of 2015. Even though this group has been active since then, this is the first group of measures that has been under this trial period even though

we're a year in. So we just want you to be aware that these are things that we asked for from the developers. The developers have been educated on this. And they are part of these measure submissions. And we've added questions into those preliminary analyses, as well.

So what are these SDS factors? They're patient characteristics that are present prior to treatment. And they are known or suspected to be a confounder of the treatment. So these are SES in one group. Socioeconomic status: income, education, occupation and employment. And then the SDS are sociodemographic factors: we're talking about language, insurance, homelessness, marital state, literacy, health literacy.

You know, I think it's important, and as staff went through all of the forms and all of the submissions, we were looking to see if these were noted in the submission forms. And we would either adjust the questions that you're asked to consider as part of that preliminary analyses. Next slide.

So during your evaluation, and this is where something has changed from your past review, you are asked to consider this list of questions. And, again, we've talked about these on the phone. And where they were applicable we left all of these questions in the forms, so none of them should be new. You should have seen them in looking at any of your forms.

But, basically, we're looking for conceptual relationships, we're looking for at least some kind of testing and analysis that was done, if someone says there was a relationship or not a relationship. And, hopefully, we would have identified where we felt something was missing, or you should consider if something is missing. Next slide.

Also, as we've talked about, and we're obviously going to be getting into this, for those of you who have made it through all of the forms, we just wanted to give you some reminders about tool-based measures.

In the past, and certainly in Phase 1

and Phase 2, we've had a lot of discussions on are there differences in the measure evaluation criteria, with specifically PRO-PMs, or the Patient-Reported Outcome Performance Measures. As staff have taken back really a lot of the information shared during these meetings, a lot of the pushback from these meetings, a lot of the discussions about interpretation of criteria from this Committee specifically, we've really been looking at the NQF criteria and how they're applicable, and are they working for different kinds of measures? Because we're seeing more PRO-PMs coming forward, we're seeing more tool-based measures.

And just for clarification, when I'm talking about tool-based measures, these are surveys, whether they're patient experience -- so, therefore, might be a PRO-PM -- or as we did last time, we had 28 functional status tools, 28 functional status measures based on various tools or instruments. Those were not PRO-PMs, some of them were clinician-based assessments, but we

really felt the rule should be the same, that we should be looking at the same type of data, the same type of analysis, the same understanding of the psychometric properties of the instruments.

So we've been having back and forth conversations internally, and then we had a conversation with the CSAC. And we're still in the process in rolling some of this out, in educating developers. So we're still dealing with criteria, but these are some notes that we want you to know that we have tried, and have been pushing the developers to provide in their submissions for this phase of work.

MEMBER KAPLAN: I don't mean to interrupt your flow there, but can you clarify that what you're -- for me a survey is a data collection method.

MS. SAMPSEL: Yes.

MEMBER KAPLAN: It's not -- so what you're talking about, really, it's like the medical record abstraction form.

MS. SAMPSEL: Yes.

MEMBER KAPLAN: You wouldn't go through the exercise for the form, you'd just go through the exercise for the measure that's represented on the form.

MS. SAMPSEL: Correct. Yeah, and I think I have a slide coming up on that. It's a really good question, because in all honesty that's one of the issues and part of the confusion that was internal to NQF as well as what are we talking about, and why is it such a big deal?

So the first reminder is we do not currently endorse surveys, tools, or instruments, that even though we have a set of measures coming in on a measure submission that might be about nursing home CAHPS, or may be about the FIM, or may be about the CARE tool, we have not endorsed that tool. Those are the data elements that feed into the actual PRO-PM.

That being said, it's really important to consider the fact that by the time something becomes a performance measure, whether it's a

PRO-PM or a tool-based measure, that tool should have been -- its reliability and validity should have been established before coming to this, and that is an NQF policy. So some of those questions are fair game but we just want to make sure the interpretation is clear: we do not endorse those tools per se. Next slide.

Again, you know, and I think folks who were here in Phase 2 will remember, there have -- we had been interpreting the NQF criteria with a higher standard for the PRO-PMs. And this especially happened in Phase 1, when Karen Pace was here, where we were constantly reminding folks -- and if you remember, a lot of data came in after the in-person meeting because folks were providing all of their data on their actual tool and no data on the performance measure.

And I think Sherrie, you brought that up repeatedly during Phase 2, and we really struggled with that as a Committee. So this is really -- and we took this to the CSAC, as well, just a reminder. You all are evaluating the

measures, and so it is fair game to say "we want the data on the performance measure." And during the algorithm where it used to almost give you an out if you had the data element or tool-based testing, then we, you know, had let measures go through before. We are evaluating the performance measures, so we want that performance measure data.

Staff went through all of these forms, we've worked with all of the developers to make sure that the performance measure data was there. Is it the right data? That's something we still have to work on. But to the same degree, we did look for it and, hopefully, every single submission should have that performance measure data in it for this go-around. Next slide.

So this is what I was talking about before. Approximately three years ago, four years ago, NQF convened a panel talking about patient-reported outcomes and what happens when you get to PRO-PM. And there are steps. And so, first of all, you have the patient-reported outcome, and,

obviously, that should be your first step or a developer's first step: what are we trying to get to?

And then you might have the PRO-M, or the patient-reported outcome measure, that's typically the tool in this process. So that's where we're talking the FIM is a PROM, the CARE is a PROM, the PHQ-9 itself is a PROM. We expect, NQF expects, that before you ever get to that PRO-PM level, that that PROM has already been tested for validity and reliability, and that it is a usable tool or instrument, so that, frankly, by the time they get to these panels we shouldn't have to be evaluating the reliability and validity of the tool. There should be some assumptions made.

At the same time, that also assumes, well, then aren't we endorsing the tool? Right now, nomenclaturally, no, we're not, but that's something that obviously -- and something we want to talk about tomorrow after we make it through everything today, and then the final measures

tomorrow, that we still think there needs to be some work done there for understanding the criteria and how they apply.

But really what you're looking at, and what we've pushed the developers for, are this last box, to make sure that the outcome performance measure, whatever it is, average change, percentage proved, et cetera, that that is what has been tested for reliability, validity, threats to validity, et cetera. And so we've been pushing really hard on that and doing a lot of education on it. Next slide.

This is just a reminder on related and competing measures. We will not look at related and competing unless the measure has been recommended for endorsement. We have, I think, a couple of measures in this portfolio that have been identified as related to other measures. And then we have the FIM measures that we'll talk about this afternoon are competing with some measures that were submitted during our last phase of work. Only if those FIM measures are

recommended for endorsement will we have those related and competing discussions. And we'll talk about that more, and there are additional materials. Next slide. Suzanne?

MS. THEBERGE: Next slide.

MEMBER THOMAS: Still a little fuzzy on the SES -- SDS did you say? Right, either. And the interplay between that and risk adjustment. Probably not an easy answer, but I think you said that by taking those into account you mask the effects of risk adjustment, or you mask risk -- I'm just not clear of the interplay between the two.

MS. SAMPSEL: So there are two kind of schools of thought when it comes to including the SES or the SDS in a risk adjustment methodology. The one school of thought is that if you include them, that you're masking the fact that there might be disparities.

MEMBER THOMAS: Okay.

MS. SAMPSEL: And then the other is that unless you risk adjust, unless you include

those, you won't know if there are disparities, or that they really are impacting the measure. And I can tell you even within NQF that are two different schools of thought. Those of us who are trained from NCQA say, no, you don't risk adjust, you don't include SDS, you stratify. And then others have a different belief.

But what we're looking for is not necessarily that there's a right or a wrong, but if somebody has included the SDS factors, if they have included a risk adjustment strategy, there's a rationale and analysis to support what they've done.

MEMBER THOMAS: Okay, thank you.

CO-CHAIR PARTRIDGE: Peter, this is -- for folks on the phone this is Lee Partridge. CSAC, during my tenure there, which ends actually this month, spent a lot of time on this issue. And it came up mostly in the context of the equity of the hospital related measures and the safety net hospitals, and was there a possibility that we were unintentionally causing harm to the

safety net network?

And the way some of us tended to think about that was the clinical adjustments for things like race, or genetic background, or whatever, are embedded in the measure already. Those risk adjustments have been done. What we're looking at is kind of should the measure itself be somehow structured differently or risk adjusted to take account of things like -- I think you'll see it here. There's one measure I remember in particular where people said -- acknowledge there was an impact if there was a different educational background on the part of the patients, but, in fact, they had tested and it didn't make any difference with respect to this measure. So, you'll see it come up.

MS. THEBERGE: Okay. So, as you folks are very familiar with our portfolio, we have about 60 endorsed measures in this person and family centered care portfolio. We had a call for measures for this project through last winter, from October to March. And we'll be looking at a

series of measures on health-related quality of life, functional status, shared decisionmaking, symptom and symptom burden, and experience of care. Next slide.

Our maintenance measure, which Sarah was just discussing, is Measure 420, the Pain Assessment Follow-Up. Next slide. The rest of the measures are new measures for endorsement in these areas.

We had a few measures that we've rescheduled maintenance review, meaning that they just were not ready for review in this phase and we will look at them at a later date. But you may have seen them in our portfolio and have wondered about them, and that's the series of CCAT measures and a health-related quality of life for COPD patients. Next slide.

We also had a number of measures that we'll be losing endorsement for several reasons. Developers have decided not to resubmit these particular measures and so they will lose endorsement and no longer be a part of the

portfolio. And we'll be having a longer discussion tomorrow, but your charge in relation to the portfolio of measures is to look for gaps, areas where we need more measures and make recommendations on where measures should be developed. We will also ask you to make recommendations on harmonization, provide input to the measure's application partnership, and make recommendations on measure concepts for harmonization. Next slide.

So just to talk briefly about the role of the Standing Committee in the measure discussion process, we ask you all to stand in as a proxy for our multi-stakeholder membership. And you work with the staff to achieve the goals of the project and evaluate measures against the criteria, respond to the comments that we will receive on this project in the next step of the work, respond to directions from the CSAC, and oversee the PFCC portfolio of measures. Next slide.

As we talked about during the

orientation call, this is the process for the evaluation that will be happening today. We've asked our developers to provide a very brief overview, two to three minutes, of their measure or set of measures, and they'll be sitting right up here at the table with us when it's their turn to have their measures discussed.

And then we'll ask the lead discussant, you've each been assigned as a lead discussant to a couple of measures, and then we'll ask you to kick off that discussion by providing a summary of the comments that were submitted beforehand and any issues that you discovered in your in-depth review on the measures. And we'd really like you to focus on areas of concern or places where there was a difference of opinion.

We'll have time for the developers to answer questions should you wish to ask them, so if you have any questions or anything you want clarified you can just ask our developer colleagues.

And as Sarah mentioned, we ask you to put up your cards when you wish to speak, and then we'll call on you. And for the folks on the phone, please do just jump in when you wish to say something and we'll put you in the queue, just let us know that you'd like to speak. And we did have a request from one of our participants on the webinar for folks to make sure to lean close to your mic so that they can hear everybody clearly.

And then we'll have the Committee discuss all the measures, then they will vote on the criteria. Next slide will get into that here.

So voting is in the order that it's presented on the measure worksheet. First, evidence, then performance gap, then reliability, then validity. Those four criteria are all must pass, so if a measure does not pass any of those criteria, it stops. Assuming it passes each of these criteria, then we'll get into usability and use, then feasibility, and then overall

suitability for endorsement.

And we know that folks often have suggestions on how measures could be improved or things they would like to see, but we do ask that you evaluate and vote on the measure specifications as they are presented to you today. Any recommendations the developer can take back and work on for future presentation, but we need to look at what's been submitted today. Next slide.

So let's talk for a moment about the voting process. You all have a clicker. It's very important that you keep track of your clicker because it's assigned to you. We will be collecting your votes via Desi over there with the laptop, so when it's time to vote you'll have to point your clicker at Desi, the computer will count up how many votes we received so we'll know if we've gotten everybody, and the computer will calculate. When you're voting the remote will briefly display your vote choice so that you can see, and you can change your vote without

duplicating, so whatever the last thing you press will be your registered vote. The folks on the phone and webinar will be submitting their votes via chat, and Staff will be submitting their -- we have special clickers for them and we'll be submitting their votes. Next slide.

So talking a little bit about what consensus actually is, a quorum is considered 66 percent of the Committee that is serving in a particular phase so we have achieved quorum. And so to be recommended, measures must have greater than 60 percent of the Committee voting yes or high plus moderate, so that means 60.001 or, you know, anything in between 40 and 60, so inclusive of both 40 and 60 is considered consensus not reached. You may also hear Staff call it the gray zone. What this means is that the Committee has not achieved consensus. There's really no decision either way, so at that point the measure continues forward. Discussion continues if it's on one of the must pass criteria, and if it's on the final recommendation the measure will go

forward to public comment, and NQF will specifically seek comments on that measure and its consensus not reached status. Anything less than 40 percent is not recommended and so it will either stop, depending on if it's one of the criteria, or it will go forward as not recommended to public comment.

I will pause here and see if there are any questions on the consensus not reached status and achieving consensus in general. No, okay. Next slide.

So just some quick ground rules. We have asked you all to have reviewed the measures beforehand, and we do hope that you will all engage in the discussion for every measure unless, of course, you are recused. And because we ask you to serve as a proxy for the NQF membership we really want input from all of our stakeholders and everyone around the table, and we do expect that there will be differences of opinion, and that's something that we look for and we want to encourage those discussions.

So as you know, we have quite a full agenda and so we ask that you try to help us keep on time; that is Staff and Co-Chairs' job to really keep us to the time but, you know, we ask everybody to help with that. And we will also do our best not to cut off any discussion that needs to happen, but just keeping in mind that we have folks coming in and dialing in. We have several developers that will be participating by phone, as well as in person.

And so with that said, I think we can actually kick off the discussion ahead of schedule, speaking of staying on time, unless anybody has any questions before we begin the measure discussion.

CO-CHAIR PARTRIDGE: Sherrie.

MEMBER KAPLAN: Just two quick things, if you could comment on the attribution issue and the purpose of measurement, and what the Committee's responsibilities are with respect to how the measure is being used, and does that change the psychometric standards that you would

use in applying the measure? And the second thing is to clarify the multi-item measures -- the items that are supposed to represent a single construct to make it clear that you're endorsing the composite of those multi-item measures, not individual items within the composite that's supposed to measure, for example, quality of shared decision making.

MS. SAMPSEL: Sherrie, it's so early. So kind of regarding attribution, and I just want to make sure I'm using it the same way that you are. You're not talking about the concept of how measures are being used with quality improvement versus accountability.

MEMBER KAPLAN: That's what I was talking about, if there's no difference in the standard, you're not -- you're supposed to be agnostic to the purpose of measurement for the purposes of this Committee versus ---

MS. SAMPSEL: Correct.

MEMBER KAPLAN: You're supposed to be thinking, for example, about what proportion of

the variation is attributable ---

MS. SAMPSEL: Correct.

MEMBER KAPLAN: -- to what you do to the patient, not who you attract in the case of the SDS adjustment, so the attribution of that --it is supposed to represent quality of performance of the unit that's being compared.

MS. SAMPSEL: Correct.

MEMBER KAPLAN: And you're not supposed to consider how it's being used, i.e., in the penalty phase of the CMS, you know, re-admissions thing, for example, or to incent physicians. We're supposed to be agnostic --

MS. SAMPSEL: That is correct.

MEMBER KAPLAN: Okay.

MS. SAMPSEL: You should be agnostic to that.

MEMBER KAPLAN: And then the multi-item thing, just to clarify what you were -- because I'm -- it's still resonating as a little bit of a confusion, the term survey versus a multi-item measure of a construct. You don't want

people looking at individual items within that multi-item composite's intended use ---

MS. SAMPSEL: That's correct. Where there's a little bit of difference for that is, for example, the second measure we're looking at today, 2967, which are 19 measures within a submission, and there's a difference between the measures that were submitted, those 19 measures, and the items that were collected on that survey. So we're not asking you to dive into each individual item on the survey, but how those are then being combined into the measure.

CO-CHAIR PARTRIDGE: Sarah, I know it's early in the morning, but I think it will probably come up fairly often. I wonder if we could just talk a tiny bit more about your statement that the -- whatever the data collection tool is, whether we're talking a survey, or as I think Sherrie said a chart abstraction, whatever, we should be comfortable that that tool has been tested. If we're not comfortable that it's been tested, we should be

raising that in the validity, reliability part of our discussion, yes?

MS. SAMPSEL: Yes, so that would be under -- I mean, technically it's under both reliability and validity because if you think about it they both have to -- they both have questions about the specifications. And so one analogy we are talking about internally is how this kind of translate -- if you have a claims-based measure, your data elements are typically some demographic data, age, gender sometimes, things like that, and then you have claims. Those are your data elements, those exact claim numbers. When you're talking about a survey or a tool, your data elements are technically the items on the survey so there is some level of having to have some comfort that the reliability and the validity of that survey, you know, is accurately performing into the measure. It's kind of hard to discern that but that is the data -- the survey and those instrument items are our data elements for these measures. So yes, those

are valid questions. But you guys have procrastinated long enough, you have to start reviewing measures now.

CO-CHAIR PARTRIDGE: Chris and I are going to tag team throughout these two days. When he's been involved particularly in working on a measure as a lead discussant, I'll take over, and vice versa. In some cases neither of us was, we'll just take turns.

I want to start by saying thank you so very much to all of the NQF Staff. Those of us for whom this is the third time around for these measures recognize the tremendous amount of work that has gone into the material that we received in preparation for this face-to-face meeting in comparison to what was available the time that we began in Phase 1. And I know it's been a lot of work for you and for all the developers and their representatives sitting behind me and sitting up here at the table, now we've pestered you with a lot of questions for additional information. You probably thought we were nuts but, particularly,

I think that the measurement -- what do we call it, the sheet that we all get to read which records through the process the comments that have come in from the public prior to our considering these measures, the questions that were raised by the Standing Committee Members, the responses from the developers. They're there and they're now all in one place that we can read in preparation for this meeting, so I hope it's going to help everything go a little more smoothly.

With that, I am going to turn to our first measure, 0420: Pain Assessment and Follow-Up. And this measure steward is CMS. The representatives here are --

DR. BERG: Yes, I'm Sven Berg. I'm the Chief Medical Officer at Quality Insights, and with me is Jane Lucas. I'll have her introduce herself, as well.

MS. LUCAS: Hi, I'm Jane Lucas. I'm the Project Manager for this project.

CO-CHAIR PARTRIDGE: Thank you. And

our lead discussants are David Cella, Linda Melillo, Rebecca Bradley, and Carin Van, I'm going to pronounce it wrong, Van Zyl. And I apologize if you all on the phone are getting a little bit of the sound of mechanics out there. There's a construction site on the other side of the street.

DR. BERG: All right. Well, thank you for the opportunity to come here this morning. It really has been beneficial to us to get many of the comments prior to this meeting, and so we appreciate the new process. Having been here a number of times before myself, I recognize that usually the first presenter gets hit with lots of questions, and so it's with a little bit of trepidation this morning that we begin the list of measures today. However, on behalf of CMS, and as the measures developer for this measure, Quality Insights is pleased to have the opportunity to discuss the maintenance of NQF 0420: Pain Assessment and Follow-Up.

This measure received initial

endorsement in 2008. It was originally developed under a contract with CMS to allow non-physician providers to report quality measures through their reporting programs which have evolved over the years. And it was first implemented in the Physician Quality Reporting Initiative, and most recently in the Physician Quality Reporting System program.

PQRS uses G-codes for the eligible professionals to document whether the quality action has been performed. The level of detail in defining the G-codes to describe the clinical actions indicates there is a relationship between the assessment of the patient on the date of encounter using a standardized tool and a care plan based on that assessment when the screening is positive.

Pain assessment is the critical first step in determining interventions, establishing goals within a care plan, and measuring the effectiveness of the care plan or treatment interventions. For example, the specification

provides the numerator quality data coding options for reporting satisfactorily. G8730, which is then a G-code is that the pain assessment documented as positive using a standardized tool and a follow-up plan is documented.

The measure specifications for NQF 0420 provide clear definition for each of six G-codes used for reporting, as well as definitions of what constitutes a follow-up plan. Measure testing by Quality Insights includes chart review to determine that the correct G-code was selected by the eligible professional, a standardized tool was utilized, and document of a follow-up plan was done when indicated.

The intent of this process measure is for an eligible professional to document that an assessment of pain has been performed using a standardized tool and that a follow-up plan, i.e., a treatment plan has been documented in the medical record. The measure focuses on the adult population, those 18 years and older, and the

denominator includes all visits occurring during the 12 months reporting period.

Since implementation the number of eligible professionals reporting the measure has increased each year, as well as the scope or the type of professionals that have reported. What I would like to point out is that although noting that the evidence surrounding the effect of pain screening on treatment prescriptions for the level of pain is of low quality at this point, the Institute for Clinical Systems Improvement included pain assessment in their most recent opioid prescribing protocol and in their guidelines for the assessment and management of chronic pain. In addition, the recently published CDC Opioid Prescribing Guidelines also include an assessment of the characteristics of pain. It notes that an assessment of pain perception over time can be a valuable measurement of pain improvement over time. Failure to assess, address, and treat pain may negatively impact the quality of life, resource use, and productivity.

A report by the Institute of Medicine entitled, "A Blueprint for Transforming Prevention, Care, Education, and Research," suggests that chronic pain rates will continue to increase as a result of Americans experiencing disease in which chronic pain is associated like diabetes, cardiovascular disease, et cetera, the increase in obesity which is associated with chronic conditions that have painful symptoms, progress in lifesaving techniques for catastrophic illnesses and injuries for people who would have previously died, surgical patients that are at risk for acute and chronic pain, and that the public itself has a better understanding of chronic pain syndromes and the new treatments associated with that. Therefore, as a result of all of these things we believe that this measure is still important to improve the quality of life, reduce unnecessary medical treatment, and undue financial burden.

We thank you for your consideration of the endorsement, and we look forward to your

questions and the discussion that is surely going to follow. Thank you.

CO-CHAIR PARTRIDGE: Thank you. David, do you want to -- David Cella, do you want to lead off?

MEMBER CELLA: Yes, hi. Can you hear me okay?

CO-CHAIR PARTRIDGE: We can hear you just fine.

MEMBER CELLA: Okay. Yes, you know, maybe I could just say a little bit up front and then have somebody in the room that was a primary reviewer like Linda or Rebecca go through the criteria. But I just -- maybe we'll start by saying this is a maintenance measure. It's the only maintenance measure we'll be reviewing, and so some of the criteria don't have to be discussed, but evidence does have to be discussed, and it's the first criterion.

The Staff review was that the data were insufficient. I actually, if we're talking about evidence, I think that there is evidence in

the literature, not specific to this measure, but in support of this measure that most of the time people study the impact of pain, you know, with monitoring and setting up a treatment plan for symptoms that are monitored, not just pain, that there is a benefit to a patient's experience of care fairly consistently. The data on benefit to actual outcomes on that symptom or general quality of life are a little more mixed but my personal view is based upon you knowing more about this than chronic disease management like cancer is that there is evidence for this practice being good for patients.

CO-CHAIR PARTRIDGE: Linda. I'm sorry, David. I don't want to cut you off.

MEMBER CELLA: Yes. Well, I really thought I should just start with a quick introduction. And we are just focusing for now on evidence. Correct?

CO-CHAIR PARTRIDGE: Correct. So which of your colleagues here in the room wants to continue? Linda? Carin.

MEMBER VAN ZYL: Yes, as a palliative care provider, pain is the number one thing that I assess, and I also had the interesting experience of trying to develop this measure for our cancer program.

In the 10 years that I've been doing this, I understand that there is a -- there's sort of a necessary but not sufficient component here; meaning you can't make pain better unless you ask about it. So the assessment is crucial, but my clinical experiences, and I think the experience -- the reason that there's so little in literature that supports assessing means better outcomes is because I don't think that there's a very good relationship between them. And, you know, the rating scales are such a poor substitute for the total experience of pain. The new medical rating scale is the most commonly used measure but at this point is relegated almost to a joke because it seems to have so little relationship to how a patient functions in the world.

I thought the evidence was insufficient here, and not on the fault of the developers. I think we really grapple with anybody studying pain trying to figure out how you put a number to it, and what that really means for somebody, and more importantly what you can do about a number that's really high, especially in chronic pain. And when we were trying to go through developing an easier way of documenting a pain assessment and a pain plan, it was a year's worth of work, the physicians were furious that they had to document this. They believed that they already were; they weren't. And the only way to go through and make sure that a physician had really sort of ended the sentence with an assessment and a plan was a manual data abstraction. So we tried to come up with a standardized note that made it easier for physicians to document, and after a year's worth of work we couldn't figure out whether there was any way of insuring that this extra step had actually meant a single thing to a patient.

I don't -- my comments are not meant as a criticism of the idea, but more a comment on the practicality as a clinician and as somebody who tried to implement this measure in the real world.

MEMBER MELILLO: So thank you for those comments. I would agree that pain is a huge issue. I would also agree it's very difficult to measure. One of my concerns is that patients frequently will gain the number scale. You know, the find out very quickly that the higher the number they give it, it changes the medication they receive. So also the FACES pain scale was designed for children, and much of the evidence was on low back issues. So I really would like more clarification on the patient populations you would use this with. Is this just for patients diagnosed with chronic pain, or if this is broader, is there any evidence of its use, any broader context that where it's been beneficial?

CO-CHAIR PARTRIDGE: We just clarify for Linda the age range. I think it's 18 and up.

DR. BERG: 18 and up, yes.

CO-CHAIR PARTRIDGE: Yes, so it wouldn't apply to children.

MEMBER BRADLEY: I love going last because -- I have --

CO-CHAIR PARTRIDGE: Just a reminder everybody down at that end of the table, we can't see you.

MEMBER BRADLEY: Yes. I had the same concerns about -- that my colleagues previously mentioned about the validity and the evidence.

I guess I was also, as I read it -- you mentioned that the number of provider -- eligible providers have increased since 2008, but participation, this is a voluntary reported measure, and participation has decreased during that time. And I was wondering if you could address that, and since it is a measure that has been out there a while, what do you attribute the decrease of participation to?

MS. SAMPSEL: Actually, I'm going to -- we have a new member who joined, so I'm going

to, Jane, give you a second to come up with an answer. And, Lisa, if you will introduce yourself, kind of a brief background, but also we need you to make any comments and disclosures of interest or conflicts of interest with any of the measures we are being reviewed.

MEMBER SUTER: Can you hear me? Okay. I'm Lisa Suter. I apologize for being late. Can you hear me? Okay. I'm an Associate Professor of Medicine and Rheumatology at Yale School of Medicine. I also work for the Center for Outcomes Research and Evaluation as a measure developer on some outcome measures under contract for CMS, none of which are under consideration by this Committee. But that's my main disclosure as a measure developer. I'm also a health services researcher. Again, I apologize for the disruption.

CO-CHAIR PARTRIDGE: I'm going to remind everybody again we are having to talk over the sound of digging machines next door. Try to remember to bring the mic a little closer to you

and that way we'll all hear throughout the room. Jane.

MS. LUCAS: I'm showing on the form that we submitted the number of eligible providers reporting the measure is about 10.7 percent, which is an increase of 3.6 percent in 2010, and then it went up to 7.4 in 2013, and 10.7 in 2014.

CO-CHAIR PARTRIDGE: Peter, you're up next.

MEMBER THOMAS: I guess I'm just having a little trouble understanding when you would apply -- two questions; when you would apply this -- does this apply to anyone who walks in for an office visit to a physician, that the intent is to do a pain assessment, and if there is pain follow-up, or is this only after a patient reports pain or chronic pain?

DR. BERG: So the measure is not specific to diagnosis, so the measure was originally developed for individuals like physical therapists, occupational therapists, et

cetera. So it was an important part of their practice, and an everyday condition for most of the patients.

It is a voluntary measure so it's a measure that a physician can select to perform. I would anticipate it would be used in practices where there is a significant amount of pain in their patients.

With regards to the other question about decreasing performance rates, I think was the question that you had. You know, when it was initially adopted then it was very specific towards the type of eligible professionals that were using it. I don't have specific information in front of me, but my -- I would suspect that the performance rates have decreased because there is a broader range of professionals who are using the measure over time, as well. But I don't have any evidence to support that.

CO-CHAIR STILLE: Any follow-up?

MEMBER THOMAS: So the materials, the Staff rated this insufficient for evidence and

the materials talk about the developer having the opportunity to provide additional information. Have you provided them additional information, or do you have something more to tell us?

DR. BERG: Well, we've provided the information that is available with regards to this. It's a fair criticism, and I recognize your comments, that it's been very difficult to take pain assessment and study whether or not there is a better outcome because the assessment has been made. And we -- after the comments were provided we did a literature research and we looked for other evidence along those lines, and really those studies have been difficult to do, and the evidence that is there is of fairly low quality on both sides. In other words, studies that have been done that look at the effectiveness of this and have found no effect have also been rated as having low quality, as well, so it's a very difficult topic to address.

We believe that because those that are continuing to develop protocols around the

management of pain continue to use pain assessment as part of those protocols, we believe that it's justified to continue to utilize this measure at this time, but we recognize the criticism that has been raised by the Committee. And our recommendation and take home back to CMS is, is that perhaps this is something that deserves additional study in the future so that we can be assured that utilization of a measure like this, which is a process measure, actually has benefit to patients in terms of outcomes. And we recognize that that is a weakness in the current science.

CO-CHAIR PARTRIDGE: Lisa.

MEMBER MORRISE: Thanks, Lee. I have a couple of questions or comments. As a mom of a kiddo who encounters pain on a really regular basis, and it's really bad, I really think this is important, because if you don't ask you don't know. And as far as the scale goes and using the FACES scale, while I agree that there's probably some issues around the scale, if I heard

correctly, Sarah, we're not looking at the scale. We just looking at the measure, so we kind of need to take our concerns around the scale and put them aside. I personally really like the Legos scale. Anyway, Lego FACES, you know. Anyway, so that's one comment.

The comment that was made by the reviewer around the fact that certain ethnicities are under assessed, I think is one reason, glaring reason why we need this measure, why we should say that it is essential because there is an under-assessment, and as a consequence of that, under-treatment of pain specifically for the African American population. And that is true in the pediatric population, as well as the adult population. It significantly impacts our friends with sickle cell disease, and we really need to ask.

The outcome piece should be paired eventually, but I think that's a separate issue also; that first you have to determine if the pain exists, and then determine if appropriate

treatment is applied, two separate issues. But if you don't ask you don't know, so those are my thoughts. I think we should say yes, it's needed.

DR. BERG: I appreciate that type of endorsement. One of the things that I would like to point out is that we're not specifying which pain assessment tool to use, so although you might find fault with the FACE scale, et cetera, that's -- we're not really -- we're not assigning a particular protocol with this. We're just asking for the use of a pain assessment tool, and so we're allowing clinical judgment to be used within each practice on how to do that.

And the other thing that's difficult because most of the study that's been done on whether or not a pain assessment has been performed is looking at just that, and then whether there's an improved outcome. This measure also asks for the addition of a treatment plan after the pain assessment, so we recognize that it was not just the assessment of pain that was important but that an appropriate treatment plan

was then provided following that assessment tool. And because this measure is looked at for each visit that a patient encounters with that provider, then further assessments of pain should be happening, we should be tracking the patient's self-reported pain over time, and as a result of that the treatment and the treatment plan can be modified in response to that pain. That's the design of this measure so we're not looking just as the assessment, but that a treatment plan has been added, as well.

CO-CHAIR PARTRIDGE: Chris.

CO-CHAIR STILLE: I think Becky had --

MEMBER CELLA: This is Dave Cella, again. I might be repeating myself a little bit, but I -- you know, I think that there is -- we're really just focusing on evidence here, not the tool, and there's not even a specific tool recommended in this particular measure. I think it's a plus that the addition of a plan to address the assessment of pain is included. I think it moves this process measure on step

closer to outcome because having a plan to management pain is also associated with improved pain.

I mean, I go back to 30 years ago, you know, when people starting wearing -- providers starting wearing pain buttons saying "Ask Me About Pain," and the No Pain button. The reason we do that is because it helps encourage patients to tell us about their pain, and then we have -- that's when you do something about it. And I'm concerned in this era of worrying about narcotics overuse and the Joint Commission pulling away from its recommendation of pain assessment that we're going to swing in the wrong direction saying hey, we've just come to realize there's no relationship between asking about pain and managing pain. And I don't think that's true. I think there is evidence. I'd like to see this measure go a step further and actually, you know, the pain outcome. And I guess I'm trusting that that's going to happen in time, but just focusing on the evidence criteria and the other issues to

discuss, I just repeat, I think the evidence is there.

CO-CHAIR PARTRIDGE: Becky, you're up.

MEMBER BRADLEY: Thank you. Several of the research articles that you cited talked about disparity, and also the -- one of the objectives is hopefully to reduce the cost and utilization of healthcare services. But as a -- a G-code is a billing code, I assume, and so this is a code that you abstract from claims. And it's kind of confusing how this would reduce cost and utilization. It seems like it might add to the cost and utilization where we have more providers assessing and documenting that they ask about pain, and then they developed a treatment plan because they can bill for those assessments now, where they could not in the past. So maybe they were asking but they weren't documenting.

I'm having a little bit of trouble getting from how it become a quality or an outcome measure and meet the goal of reducing costs versus increasing costs.

DR. BERG: And, again, this is not -- it's not an outcome measure, and so although we would like our processes to be associated with good outcomes, it's not what the measure was designed to do. Our thought process here is that the appropriate assessment of pain and then appropriate treatment plan will result in less pain, and the treatment of those conditions then over time should then decrease healthcare utilization, and so that's the theory. Again, we're looking for that connection and that evidence. And that evidence of what the outcome is from this we agree needs to be stronger than it is right now.

CO-CHAIR STILLE: As a pediatrician, I just have to ask this question. Given that there are good tools for measuring pain in patients of all ages, what rationale do you have for excluding patients under 18?

DR. BERG: Can you say what that --

CO-CHAIR STILLE: Well, again, I think it -- the measure, when the measure was developed

was -- that was the population that was being used, and so it hasn't been updated from that time.

DR. BERG: So just had to get out there.

CO-CHAIR STILLE: Sure.

DR. BERG: I'm a pediatrician, as well, so I understand where you're coming from.

CO-CHAIR PARTRIDGE: Are there questions from any other Committee Members in the room? I'm sorry, Nicole, I'm just looking right past your card. Go ahead.

MEMBER FRIEDMAN: I have two quick questions. My first one is in terms of dropping the G-code or the code, is that something a non-licensed provider can do, like a community health worker? And is there evidence that increasing the care plan reduces non-pharmacological intervention to improve pain?

DR. BERG: We would hope that the care plan would be appropriate for the patient, and so if non-pharmacological intervention was the

intervention that should be used, then we would hope that that would be the care plan. Again, we're not specifying what the treatment is, we're allowing for the clinician to use their best judgment in determining what that treatment plan should be. So this is not pain assessed, provided opiates or other pain medication. That's not the purpose of this measure.

The first question, give me a little -- remind me again what that was?

MEMBER FRIEDMAN: In the documentation part of the --

DR. BERG: Sure, okay, whether or not non-physicians and community health workers can do that. I don't know the answer to that question. I don't know exactly who can -- who's allowed to build specific G-codes. I don't know if anyone else on our team knows that.

MS. LUCAS: It is -- CMS defines who the eligible providers are, so I don't have that list in front of me but it would include like nurse practitioners, chiropractic, social

workers, psychologists, et cetera. But there is a list that's out there.

DR. BERG: But you specifically addressed community health workers, and I don't know the answer to that question.

DR. GREEN: So this is -- can you guys hear me?

CO-CHAIR STILLE: Yes.

DR. GREEN: This is Dan Green from CMS. So the G-code just to be clear is only to add a quality data code, I believe, in this measure, to indicate that the quality action was performed or not performed. I don't know if that helps at all.

CO-CHAIR PARTRIDGE: Peter.

MEMBER THOMAS: In other words, it's not a billing code. You don't get paid under a G-code, it's just merely a data collection code?

DR. GREEN: Generally speaking, G-code -- yes, generally speaking, G-codes are not reimbursable. There are some G-codes for certain services, I want to say Pap smears, for example, that may be associated with reimbursement. But

generally speaking, the G-codes that we have in our program are for quality reporting, just quality data codes where there's not a CPT II code, you know, developed by the AMA, for example.

CO-CHAIR PARTRIDGE: Peter.

MEMBER THOMAS: So given the importance of pain management and the opioid epidemic, I mean, this is a very important area, so I'm really tossing and turning on this issue of insufficient evidence. And I guess what I'm wondering is -- I'm trying to establish a -- what's the threshold for this? So in the real practical scenario where a patient comes in to a physician and the physician asks kind of oh -- even just kind of routine questioning are you -- you know, do you have pain? And the patient says well, I've got a little low back pain but I'm fine, and they move on in the conversation. Is that an assessment? Does that qualify to the level of an assessment, or is that not really what you're talking about? You're talking a more

formal thing.

DR. BERG: When we do the chart review we are looking to confirm that the G-code actually matched what happened. That's part of the validation that we'll talk about later. We are looking for a specific pain assessment tool.

MEMBER THOMAS: Okay.

DR. BERG: Not just a --

MEMBER THOMAS: Not just a random --

DR. BERG: Not just a question about pain.

MEMBER THOMAS: It's higher level than that. Thanks.

DR. BERG: Yes.

CO-CHAIR PARTRIDGE: Are there more questions from the room? Jennifer.

MEMBER BRIGHT: Sorry, I didn't do my tent thing first. Okay, I'm a novice at this, so bear with me while I ask what might be an obvious question. But listening to all this, I certainly agree with Lisa that there is huge value in assessing, asking patients, especially those that

are under-represented what their level of pain is. And I certainly agree with the marrying of assessment with a plan. Too often we -- those are disjoint. But I guess when I'm thinking about this, what we're talking about is this is a measure of quality, and what I'm hearing is that we don't have evidence that we can associate this process with actual quality of care. So we know through this process measure the data that you've collected for the last six years or so tells us that there are people doing this, and that it's happening the way it should be happening. But what we don't have evidence for is any indication that it's resulting in quality of care. We're interpreting that. So I guess where I come down on this is, why are doing the -- why is this measure valuable then? It's not to say that pain assessment is not valuable, it's not to say that we shouldn't be doing it better, but what I'm hearing is that we don't have evidence to tell us that this measure is making an impact, or we don't know yet, there needs to be more study.

So in my job in this Committee as I understand it, I guess my question is why? Why re-endorse this measure if what we're really saying is we need a better measure?

DR. BERG: Again, I think it's a fair question. Is that Dr. Green? You can -- I'll start then.

In many situations when research studies are designed and you're going to do a controlled study between one arm or another arm, you have to determine whether harm is going to happen to your patients and what the risk of each of those study arms are. This is a very difficult topic around which to develop a controlled study because obtaining a patient history, and developing then a treatment plan around a history is the standard of care. And to design a study that would exclude that so that you can prove that this is associated with a good outcome might be an unethical study, so this is a very difficult study to develop. About the best you could do is do some type of retrospective review

comparing practices that utilize this tool with practices that don't utilize this tool. Those studies have not been done, but ethically this is a very difficult study that you're asking in order to obtain the evidence that you're asking for.

MEMBER CELLA: This is Dave Cella, again. I might be guilty of not understanding the evidence criteria. I may be being too broad in my interpretation of it, but although I accept that this particular specific process measure does not have an evidence base in the literature, there is a good size evidence base in support of the practice of monitoring symptoms and altering practice based upon that monitoring and the effect that has upon patients. I mean, there have been enough studies now where there have been reviews of the study that I'm just talking about in oncology and palliative care where, as I mentioned earlier, fairly consistently satisfaction with care has improved by this practice, and less consistently but with a small

effect overall symptoms are improved with this practice in well controlled randomized clinical trials. So that's -- but not with this particular performance measure, so if the evidence we're supposed to review is only on this measure or on the practice of monitoring symptoms and making treatment plans based upon that monitoring practice, I'm suggesting that the evidence is there. And what's not clear to me is whether that is part of what we should be considering as evidence.

CO-CHAIR PARTRIDGE: Jennifer, does that help you?

MEMBER BRIGHT: It helps me, but it doesn't convince me.

CO-CHAIR PARTRIDGE: Fair enough.

MEMBER BRIGHT: I guess my take on this is -- a follow-up question is would this practice continue? It's a voluntary measure, number one, so it's not required, but is your sense that the measure is required to keep the focus on this becoming standard of care, because to me that's

the purpose of a quality measure, is to both incentivize and insure that there's a consensus understanding of what to do to lead to quality. But if this is (a) a voluntary measure and (b) not really demonstrating correlation with quality outcomes, how are we accomplishing that?

DR. BERG: So, again, it's -- further in the discussion, if we get past the initial vote. But, again, speaking to the gap that exists, there is a practice gap. And it was noted that the performance as more providers have been utilizing the measure, the actual performance has been diminished, and so based on that performance gap we believe that there's still a place for this measure because it's not being utilized by all of the eligible providers that are actually reporting the measure.

CO-CHAIR PARTRIDGE: Sherrie.

DR. GREEN: So this is Dan Green, also. I mean, I know we're not talking program-specific, but for parts of Medicare, I mean, we have over 300 measures in our Ambulatory Quality

Reporting Program. We don't require really any specific measures. We do require some specific types of measures in the program. Obviously, I'm sure there are other quality reporting programs around, but I think we're probably the biggest. So I understand your point about it not being a mandatory measure, but just for context, none of our measures specifically are mandatory because we can't tell a provider, you know, a particular measure is or is not inherent to the care that they give.

CO-CHAIR PARTRIDGE: I'm going to go next to Lisa on the phone because she's been waiting. Liz, excuse me. Too many L's, aren't there?

MEMBER MORT: Thank you, Lee. I just wanted to add into the conversation about the evidence. And I feel that in addition to not knowing whether assessment of pain and an action plan leads to outcome, it's worth noting that -- and when I read over the detail that the specifications for what the action plan are are

very broad. It really just means you're looking for the provider to document that they either gave medication or made a referral. And I think that we're really in the throes of trying to understand how to manage pain better, you know, in light of the opioid epidemic, Fentanyl, et cetera. We're all getting lots and lots of direction, not all evidence-based but we're getting a lot of direction to back off on narcotic opioid prescriptions and use other non-pharmacologic or non-opioid prescriptions. So in light of the change in the Joint Commission, in light of what's going on nationally, I feel that there's too much uncertainty in the best approach to pain management to endorse the measure at this point in terms of the evidence.

CO-CHAIR PARTRIDGE: Sherrie.

MEMBER KAPLAN: I think we're getting a little bit confused between the two competing -- I'm hearing two competing kind of issues. One is the under-detection of pain and suffering, and the ability of an assessment. And the old wheez

is in public health terms if you monitor it, you'll detect more of it. So, you know, there's the sort of is pain really being under-evaluated and, therefore, under-detected and unnecessary suffering is going on versus the use and usability issue which is different from the evidence-base issue of are there going to be unintended consequences of implementing this measure in terms of elevating the prescribed -- you know, the number of prescribed opioids, et cetera, et cetera?

So, Sarah, help. Are we drifting into use and usability here, or are we -- is there -- what's being -- what evidence are we looking for?

MS. SAMPSEL: That's a great question. I mean, yes, I think there have been some other questions that have been trending down the use and usability side, and specifically we're going to talk about that in that section of the measure. But, you know, what I can talk about is the Staff review and the Staff analysis, and to

interpret why we gave it as insufficient. And that was because of the fact that we weren't able, and we were looking specifically for clinical guidelines, USPSTF recommendation, or the empirical evidence citing that the actual documentation of a pain assessment and actual documentation of a care plan led to, you know, a change in the intervention or improvement in the intervention which then impacted the outcomes, and we weren't able to find that. But that's not inconsistent with other pain measures, and other measures even, you know, kind of how David Cella has been talking about it, that there are some things that, you know, it's kind of if you don't measure it, how are you going to know if it's happening or not? And, actually, the Palliative Care Committee reviewed a very similar measure to this, as well, and they went down the line of evidence insufficient but how do you argue against the fact that assessing for pain and putting a treatment plan in place is not important? And, therefore, that's how they

passed, it was insufficient evidence with exception, which is available here. So, I do think it's been going down both paths, and what I would suggest you all think about is, you know, first of all the requirement for the process measure is some kind of guideline or empirical evidence saying it's a good thing. If not, you have the exception possibility, and really kind of -- you know, and I think Jennifer got to the question, as well, is then, you know, kind of what does this do with standard of care, and is there evidence that supports it? And is there the need for the quality measure, and that's what we're looking for under importance and evidence.

CO-CHAIR PARTRIDGE: I saw a card up down at the end of the table. Did it go down? No, all right.

MEMBER AVERBECK: A question. You've had about 10 percent it looks like of eligible providers submitting. If the goal is around a screening measure for pain assessment, I'm just curious because it does require a G-code. Are you

finding that the types of practices that are using it for the intended use is around sort of general practices, as opposed to practices that might be dealing more with oncology patients, palliative care? So just wondering if your field experience is reflective of what the goal is, which is to move it towards population screening.

CO-CHAIR PARTRIDGE: Jane, I'm not sure you have the data to answer that, but --

DR. BERG: She's furiously looking to see whether we can answer that question specifically.

MS. LUCAS: It looks like in 2014 we have the number of claims submitted by classification of eligible professionals, and it looks like about 50 percent were chiropractic, and about 12 percent were rehab and restorative providers. One of the issues that we have with trying to drill down into the types of providers is that there's a pretty large list of providers that we have, and so we're not always sure which bucket they will fit into, but pretty much it

looks like it's physical therapy and chiropractic in 2014. And that was the same year that we actually expanded some of the denominator codes for others to report.

CO-CHAIR PARTRIDGE: All right. If there are no further questions, I think -- am I right? I'm sorry, Linda and Lisa.

MEMBER MORRISE: I was just going to suggest that we -- because I'm hearing mixed responses, that we go ahead with the concept of no, but exception because I think what we're looking at here is that there may not be hard evidence, but we do know that if you don't ask you're never going to get -- I mean, it's the first step toward collecting that evidence, and so if you take that away then we're not going to get those interventions.

The other thing I point out, too, is kind of what was just brought up, is that this isn't just for physician practices, but was meant for expanded use across those who may be assessing pain, so physical therapists,

chiropractors, et cetera.

CO-CHAIR PARTRIDGE: Linda?

MEMBER MELILLO: Yes. It's my understanding that physical medicine and rehabilitation is one of the few specialties that's not required to submit quality data. Is -- does anybody -- is CMS still on the phone? I believe that's the case, so how would you go about -- if -- I mean, you would want to include those folks, including the outpatient physical therapy, so how would you get data from those populations if they're not reporting quality measures?

DR. GREEN: This is Dan Green, again, from CMS. PM&R are not excluded from PQRS, for example, and physicians practicing in physical medicine and rehab, at least as proposed under MACRA and the MIPS program will also not be excluded from participation. In fact, they have to participate or suffer reimbursement adjustment.

CO-CHAIR PARTRIDGE: I am watching the

clock a little bit. We still have more work to do on this measure before we conclude, but I think it is time to take our first vote. And Sarah is going to tell me exactly what we're voting on so you all are clear.

MS. SAMPSEL: Yes, I will do that. Suzanne just took a drink of water. So the first thing we are going to vote on is the evidence, and should -- we'll see how these votes come out, but should the majority fall under insufficient is when we'll talk about the exception, and we'll let the voting tell us which way to go. So, basically, what you're voting on now will be evidence and, Desi, are you ready to take over?

MS. QUINNONEZ: Yes. So we are going to be voting on Measure 0420. For those who are on the phone the first option is high, second option is moderate, three is low, and four is insufficient. So option one is high, two is moderate, three is low, and four is insufficient. So voting is now open for evidence.

MS. THEBERGE: Just a reminder for our

Committee Members on the phone, please submit your votes via the chat. Thank you.

MS. QUINNONEZ: We're looking for 21 votes so we're almost there. Okay, voting -- all votes are in and voting is now closed. So for option one we have zero percent who voted for high, 33 percent for moderate, 24 percent for low, and 43 percent for insufficient. So this measure fails for evidence.

MS. SAMPSEL: Sorry, we've never seen numbers like this before. Hold on a minute. So the measure did not fail. Basically, what we would do is take the moderate plus the insufficient and move forward, and we do think we should vote on the evidence with exception because of the high percentage under insufficient and continue the conversation, but the reality is consensus was not reached on this so we'll continue the discussion, move forward and there'll be an opportunity for reconsideration and revoting after public comment. But let's move forward to evidence with exception just so we

know what that result would be.

MS. QUINNONEZ: Okay. We will be voting for evidence with exception for Measure 0420. Option one is insufficient evidence with exception, and option two is no exception. You can submit your votes.

MEMBER SALIBA: This is Deb. Can you please explain what no exceptions means in this particular vote?

CO-CHAIR PARTRIDGE: It's no exceptions. It's with exception.

MEMBER SALIBA: I'm sorry. So the first option is -- it's not showing up on my screen what the options are.

MS. QUINNONEZ: Yes, I'll repeat. I'm sorry. Option one is insufficient evidence with exception, and option two is no exception.

MEMBER SALIBA: Yes, I was asking what -- if you could give me some definitions of what number two, no exception, means.

MS. SAMPSEL: It basically means that you do not feel that the measure should move

forward, so it would fail on evidence. So insufficient evidence with exception means that you -- you know, there may not be enough evidence available to support the exact measure construct but you feel it's an important enough measure despite the lack of evidence. The no exception means you don't agree that it doesn't matter what you do with providing more evidence, you still don't believe that the measure should be supported based on no evidence.

MEMBER SALIBA: Thank you very much, that's helpful. Thank you.

MS. THEBERGE: Deb, we still need a vote from you. Thank you.

MEMBER SALIBA: Did you get the vote?

MS. THEBERGE: Yes.

MEMBER SALIBA: Okay.

MS. THEBERGE: Can everyone vote one more time?

MS. QUINNONEZ: There we are, we have it now. Voting is now closed for evidence with exception for Measure 0420. The results are 90

percent voted yes for insufficient evidence with exception, and 10 percent voted for no exception. Voting is now closed for the evidence with exception for Measure 0420.

CO-CHAIR PARTRIDGE: Moving on, Sarah, you have to help me tee up the next part of our discussion.

MS. SAMPSEL: Sure, we'll just move to the next set of criteria and so that if -- and, Kirsten, if you want to scroll down on the measure information form now, going into the performance gap and opportunity for improvement, and starting with the lead discussants, if you had any comments about that. And we can start thinking about has it already been said, is there anything you want to -- because I think we crossed over into this a little bit, but are there any remaining comments about performance gap, because there have been some discussions about disparities and percentage of providers reporting, stuff like that.

MEMBER CELLA: This is Dave Cella, just

to say I agree with the review, and have nothing to add. There are disparities on this measure.

CO-CHAIR PARTRIDGE: Any of our other lead discussants want to add to David's comment?

MEMBER MELILLO: Yes, this is Linda.

CO-CHAIR PARTRIDGE: Linda.

MEMBER MELILLO: My only comment is that from a patient-centric perspective outcome would be really important, as well. So I know that there's a process when you have to reporting first, but just to keep that in mind.

CO-CHAIR PARTRIDGE: Any further comments? Okay, then we can proceed to vote.

MS. QUINNONEZ: Voting is now open for performance gap for Measure 0420. Option one is high, option two is moderate, option three is low, and option four is insufficient. All votes are in and voting is now closed for performance gap for Measure 0420. We have 43 percent high, 57 percent moderate, zero percent low, and zero percent insufficient.

MS. SAMPSEL: So we'll go ahead and

move forward and back to the lead discussants to start talking about scientific acceptability, specifically reliability. And the first question there is about specifications.

MEMBER CELLA: This is Dave Cella. I think -- is this a category that doesn't require a vote necessarily, because it's a maintenance measure?

MS. SAMPSEL: So technically, Dave, yes, except for the specifications have changed considerably over time. The first --

MEMBER CELLA: Okay. Well, I was just -- I would just say, and to repeat the conversation earlier, with regard to the PRO-PM that is the performance measure, reliability is very high. This is not about whether pain is reliably measured itself but whether the pain performance measure -- the performance measure of pain assessment and treatment plan was documented. The reliability is quite high, so I would give it a pass.

CO-CHAIR PARTRIDGE: Any other

comments? All right, ready?

MS. SAMPSEL: Desi, are you ready to move forward?

MEMBER KAPLAN: Are we voting on the whole reliability issue or just the specs part?

MS. SAMPSEL: We vote on the whole reliability, so if there are any additional comments about the testing of the measure as it's been retested?

CO-CHAIR PARTRIDGE: Sherrie.

MEMBER KAPLAN: You have to help me with this because this may be one of those issues where the rules may be about to change and are not applicable in this circumstance. But the between provider variance doesn't look very big, and if you -- I don't know if you actually tested the reliability at the provider level because the unit being compared listed is the clinician or the group level, but the reliability data provided were kappa agreements and it looks like patient level. So can you help us understand what between -- is that a fair issue, or not?

MS. SAMPSEL: Yes. No, they should talk about their --

MEMBER KAPLAN: Okay. So if it's being used at the provider level, what did you do to test the reliability at the provider level like intraclass correlation coefficients? What proportion of the variance -- you provided something here that says that only 10 percent of the variation is attributable to the provider.

DR. BERG: Gary, can you answer that question?

MR. REZEK: Yes, I don't think that the percent of the variance attributable to the provider was not something we provided, I don't believe. The -- so the performance score reliability, the signal to noise data was based on all of the claims and registry data. We tested each separately reported. And, again, I'm not quite clear on what the question is, but we also did -- so the kappa results are basically a sample of that data where we extracted data from the medical record between two independent

reviewers to basically test repeatability of the measure spec. So the inter-rater reliability with the kappa is essentially a different analysis than the performance measure score reliability.

MEMBER KAPLAN: So on page 41 there's a column listed between provider variance, and it suggests that there's only like 10.5 percent of the variance is attributable to providers so I wasn't clear what that meant. Did you do the -- did you test this at the provider level versus just the agreement between two data sources?

MR. REZEK: Yes, that is at -- that is variance at the provider level between providers defined by basically NPI.

MEMBER KAPLAN: So just to make sure I understand, 10.5 percent of the variance in the claims data, for example, was attributable to the between provider variance.

MR. REZEK: I think that that's a fair interpretation.

MEMBER MELILLO: So this is Linda. On one of the submission pages, I have the form,

page 3 where you note performance variation, showing that 90.6 percent are in the 25th percentile, and there is nobody lower than that. Just that there's a very high floor for this measure, and I was wondering how you interpret that?

MR. REZEK: Yes, I would say the performance scores are sort of highly skewed towards high performance. Again, I think where we sort of have difficulty making -- the problem is we have only 10 percent of eligible providers reporting so it's kind of difficult to generalize that, you know, that high performance across the entire population.

MEMBER MELILLO: Okay, so the majority of those who are reporting are doing well, obviously. Do you feel that the ones that are not reporting perhaps are not doing as well?

MR. REZEK: That is certainly not something that I would --

MEMBER MELILLO: Okay.

MR. REZEK: -- that I have information

to --

(Simultaneous speaking)

MEMBER MELILLO: It just struck me there was a very high --

CO-CHAIR PARTRIDGE: I think that's always an issue with a voluntary measure.

MEMBER MELILLO: Yes.

DR. GREEN: Well, that's exactly right. This is -- from our experience with PQRS we can't say with certainty that the folks that aren't reporting are not performing as well but, obviously, people tend to pick the measures that they do perform more highly on. When there's a small percentage of folks reporting, we can't conclude that the measure is tapped out because the argument is well, what about the other 80 percent of folks who are not reporting, for example? So we can't say with certainty your assumption, but that's some of the general feeling behind them.

MEMBER MELILLO: Yes. Okay, thank you.

CO-CHAIR PARTRIDGE: Are we ready to

vote?

MS. QUINNONEZ: All right. Voting is now open for reliability. Oh, one comment?

MEMBER CELLA: No, okay, it's okay.

MS. QUINNONEZ: Voting is now open for reliability of Measure 0420. First option is high, second option is moderate, three is low, four is insufficient. First option high, second option moderate, third option low, and four for insufficient. Okay. All the votes are in and voting is now closed. For reliability of Measure 0420, 24 percent voted high, 67 percent voted moderate, 5 percent voted low, and 5 percent voted insufficient. Measure 0420 for reliability passes the reliability criterion.

MEMBER CELLA: Okay, for validity?

CO-CHAIR PARTRIDGE: Yes, on to validity.

MEMBER CELLA: Lee, that was me that caused you to wonder if there was still someone just wanting to say something. This is Dave Cella, again. I think that the latter half of

that reliability conversation is actually a relevant validity conversation, and I think the validity is a little less than high, a little less than clear, also partly related to the original evidence conversation.

I want to just maybe clarify that in this case what we're talking about with regard to validity is the validity of that measure as a way of differentiating physicians with practices, clinicians with practices on equality and not necessarily our earlier conversation about the validity of the measure in terms of improving care. Is that right? I know that's -- that might sound like a non-distinction, but --

CO-CHAIR PARTRIDGE: Chris is nodding his head, David.

CO-CHAIR STILLE: Yes. David, I get what you're saying, and I think that's a very important point.

MEMBER CELLA: Okay. So, you know, along the lines of, I think, what Sherrie was pointing out and asking about in the

conversation, if you go to 2b5 on the meaningful difference, you'll notice that this 25th percentile even though the mean is 82 percent, the 25th percentile is 91 percent if you average up, round up, so there's a -- it's a voluntary program of a small percentage of eligible people. And, apparently, there's a small group of people that are probably doing really badly to pull that mean down from a median of 100 percent because there's some people that are participating and just making sure they do it all the time, in fact more than half. There's a group that's participating and maybe they don't realize they're participating and maybe not documenting at all. So I guess it's a matter in terms of the way I'm thinking of it is, you know, do you look at this glass as half full or half empty? Within this narrow band of voluntary participants there's a subgroup of something less than 25 percent, maybe only 10 percent given the distribution statistics we have that are really not doing very well at all. And then the question

is, is that distribution which was probably contributing to the good reliability but low coefficient that Sherrie pointed out. Is that enough to be not considered a valid measure? And I think that may be part of why the Staff review came down moderate at its evaluation. That's probably where I land, too, although it's kind of moderate tipping to low in my mind. Welcome other discussion.

CO-CHAIR PARTRIDGE: Further discussion?

MEMBER CELLA: Oh, maybe just one other thing. I'm sorry, Lee. Another thing that I didn't mention that's factoring into that is that low response rate. You know, we don't know about 89 percent of eligible people. I don't know how that plays out as a voluntary measure. That just means you do it if you're -- if you want to include it in your panel of performance measures.

CO-CHAIR PARTRIDGE: Chris.

CO-CHAIR STILLE: This is kind of a shot in the dark but apropos of David's question.

You don't have a visual representation of the distribution of responses at all?

DR. BERG: Do we? I don't believe that we've prepared one in that way. We could provide it for the Committee as a follow-on, but we don't have that right now.

MEMBER CELLA: But you do see -- I'm sorry to interrupt, but you do see that the 10th percentile is zero so, you know, if there's essentially not a lot of people between zero and -- between 1 and 90, so it's just a pretty good group of people that are just not doing it at all.

CO-CHAIR PARTRIDGE: Sherrie.

MEMBER KAPLAN: I guess to follow-up on David's issue, the issue -- if this is going to be used at the clinician level, then some demonstration that it actually discriminates practices or clinicians one from another would really be helpful. And if you've got a skewed distribution problem that's going to be an issue, and it's going to be tough in a voluntary

measure. So in fairness to the measures developers, that kind of skew in distribution you're really going to see meaningful -- you may be able to demonstrate meaningful differences but they're going to be -- or differences but they're going to be so small they wouldn't necessarily even be meaningful. So if this is only going to find those few outliers that really have, you know, the -- that really are providing crummy care, as mentioned by your current -- the way it's currently being developed, you know, I think we're kind of struggling with if only this had a better distribution, but what we're staring at is the discriminative, the discriminate validity isn't really optimal.

DR. BERG: And again, as it's been pointed out it is a voluntary measure, and so those who are reporting the measure are most likely going to be concentrating on that. I guess one could draw a certain sense of satisfaction from the idea that those who are measuring it are concentrated on using that measure, as well. But,

again, there are -- as has been pointed out there's a small portion that are not performing well that causes that, you know, overall -- the overall performance rate, you know, amongst all that are using it is still down around 82 percent, and so there is room for improvement there.

MEMBER BRADLEY: I had a question, and I guess it relates to the performance and the voluntary participation because there are other measures of pain that other -- that clinicians are using and reporting in addition to this one. So is this one -- do you envision this being on top of or in place of, or -- and does that speak to why you have such low participation because perhaps they're using other measures to report --

MS. SOMPLASKY: This is Nancy Somplasky, and I apologize, I was unable to get through on a speaking line until a few minutes ago. I'm the Director for this project, so I did want to offer up some background to help folks

understand both the history of this measure, but why the importance of it.

The PQRS program allows for non-physicians, physical therapists, occupational therapists, psychologists, social workers to report. This was originally intended as a non-MD/DO, non-physician type measure so that those physical therapists and occupational therapists would have a measure to be able to report. It was so -- it was well received and thought to be important enough that we did have a request from chiropractors to expand to include those MD/DO measures, but it is a claims and registry based measure only. So folks who -- which many folks now with the adoption of electronic health records are reporting electronic pain measures or using their EHR to report their PQRS, their voluntary PQRS reporting program. These measures are only claims or registry based, so what we find is that if a physical therapist is not linked to the Medicare reporting, they are not able to -- they may be measuring the pain but

they're not able to report that. And that's been an important distinction that we've heard that within certain health systems they may not be the ones who actually have the billing that submission go in because this is submitted via a claims form into CMS for the program. It's a little bit different than what you might see in a hospital for electronic reporting, and certainly from a physician practice where I think the mind set tends to be for a lot of the measures being reported.

CO-CHAIR PARTRIDGE: Thank you. I hope we're ready to proceed to vote on validity. Desi?

MS. QUINNONEZ: Yes. Voting is now open for validity of Measure 0420. Option one is high, option two is moderate, option three is low, and option four is insufficient.

MS. THEBERGE: Deb, we still need your vote.

MS. QUINNONEZ: Okay, voting is now closed for the validity of Measure 0420. We have 10 percent voted high, 52 percent voted moderate,

29 percent voted low, and 10 percent voted insufficient. Measure 0420 for validity has entered the gray zone. Now it passes, this measure passes.

MEMBER CELLA: Feasibility. This is -- feasibility is high. It's administrative data primarily. It's been in use.

CO-CHAIR PARTRIDGE: Looks to me as though there are no further comments so, Desi, I think you're up again.

MS. QUINNONEZ: All right. We are voting for feasibility for Measure 0420. Option one is high, option two is moderate, option three is low, and option four is insufficient. Okay, we're waiting for one more vote. Voting is now closed for feasibility of Measure 0420. Option one, 67 percent voted high, 29 percent voted moderate, 5 percent voted low, and zero percent voted insufficient. For feasibility Measure 0420, feasibility.

CO-CHAIR PARTRIDGE: Usability and use. David, any comments?

MEMBER CELLA: Yes, it's also high, although, I would add that we probably should at least mention the potential unintended consequence of narcotic overuse, although I think the benefits of the measure outweigh the potential. There's no data that that's the case, but it's on everyone's mind so I thought I would mention it. I think the benefits outweigh the potential.

CO-CHAIR PARTRIDGE: Sherrie.

MEMBER KAPLAN: Sarah, can you -- is there an opportunity to put some comment in from the Committee about monitoring unintended --

CO-CHAIR PARTRIDGE: Yes.

MEMBER KAPLAN: -- consequences of the use in case, you know --

CO-CHAIR PARTRIDGE: Absolutely. I think, actually, as the summary of this discussion this morning is written up you will clearly see that reflected because it came up right in the very beginning. Any other -- yes, I'm sorry. Lisa, right in front of me.

MEMBER MORRISE: I'm just going to say that while asking about pain is not the same and developing the care plan is not the same as overuse per se, I would point out that in the case of many individuals I know with chronic complex conditions, they actually tend to under-report their pain because they are so used to it. So while some people may game the system as Linda suggested to overreport their pain because they are seeking, there are others who don't want intervention, who do the same thing going the other direction, and that's a different study. So I think in terms of the usability, it's easy to use. You ask the question, you develop the plan. So that's my insight.

MEMBER CELLA: Agree.

MS. QUINNONEZ: Voting is now open for usability and use of Measure 0420. Option one is high, option two is moderate, option three is low, and option four is insufficient information. We're looking for three more votes. Okay, all votes are in and voting is now closed. For

usability and use of Measure 0420, 48 percent voted high, 52 percent voted moderate, zero percent voted low, and zero percent voted insufficient information. Usability and use passes for Measure 0420.

CO-CHAIR PARTRIDGE: Before we take the final vote on this measure are there any other comments either from David and company on the phone or here in the room? If not ---

MEMBER CELLA: None here.

CO-CHAIR PARTRIDGE: Okay. Sarah?

MS. QUINNONEZ: Okay. Voting is now open for the recommendation for overall suitability for endorsement for Measure 0420. Option one is yes, option two is no.

CO-CHAIR PARTRIDGE: I think I'm confused. Could we go back to the evidence insufficient with exception issue? If we didn't reach consensus on that point, but we passed everything else, we can now vote yes or no up or down for recommending to go forward. Is that correct?

MS. SAMPSEL: That's correct.

CO-CHAIR PARTRIDGE: Okay.

MS. QUINNONEZ: Okay, so we're still looking for votes for overall suitability for recommendation for endorsement for Measure 0420.

MS. THEBERGE: Liz, we need a yes or a no? Thank you.

MS. QUINNONEZ: All votes are in, and voting is now closed; 95 percent voted yes, 5 percent voted no. For Measure 0420, this measure has passed the usability and use criterion. Excuse me, this measure has passed for recommendation for endorsement.

MS. SAMPSEL: So just a couple of comments here. So, first of all, you know, that first vote about, as Lee brought up about the evidence where consensus was not reached, but then there was the pass with exception. What we're going to do there is during -- accept public comment when this measure goes out for public comment, and should anything come up in public comment you'll have an opportunity to re-

discuss in your post-comment call and reconsider if anything is changed. In the same mechanism the developers have an opportunity if they want to pull together more evidence, provide that to you during that public comment section, as well, for re-discussion to see if that vote changes.

We've done this in the past where you all, you know, get all the information and you make the decision do you want to re-vote or no re-vote, but I think we have a clear indication, or a pretty clear indication that overall this measure is suitable for endorsement. There are some kind of open questions on how to translate that evidence piece.

The other thing we just wanted to mention here and kind of because that area was so questionable, we would typically have a related or competing conversation, but I think -- and we would draw your attention to the related and competing section of the measure evaluation form where -- and the original measure information form that Quality Insights presented. There are a

number of measures that are identified as related, but Quality Insights also provided a lot of information on why those measures are different, and so would technically be related measures.

If you have any comments now about opportunities for harmonization, the overall need for measurement on pain, the overall need for measure developers to be working together to be measuring pain and documenting correctly, I think we've pulled out a lot of those in the discussion, the overuse issues, potential under-use issues, and how to translate those, but this would be the time where if you have any final comments about what the developers should think about related to pain measures, this is where we would have them.

MEMBER CELLA: Can I make a comment there?

CO-CHAIR PARTRIDGE: Yes, David, go ahead.

MEMBER CELLA: This is Dave. It's

really an encouraging or maybe even a pressuring to move to pain as an outcome, as a reported outcome measure. This is a process measure, and I think a few people, I being one of them, you know, have commented on that along the way but I don't think it came out as an explicit, I guess, position or perspective of the Committee. I don't know if others agree with that, but I'd like to see it move in that direction.

CO-CHAIR PARTRIDGE: Linda.

MEMBER MELILLO: Yes, just that a lot of the patient satisfaction surveys include questions on pain, and so that might be an opportunity to see if you can match those numbers, correlate. Thank you.

CO-CHAIR PARTRIDGE: Becky.

MEMBER BRADLEY: Just to clarify that last conversation from the person from CMS that gave us the history of this measure, so it's currently only being reported in the physician quality reporting as a voluntary measure but it was intended for other providers to be able to

report quality measures, so I'm a little confused now as to the sample that you selected to review. Is it from the Physician Quality Reporting Program only, or was it from the other providers that are submitting claims for this?

DR. BERG: So the PQRS Program has expanded to include a number of eligible professionals now.

MEMBER BRADLEY: Okay.

DR. BERG: So it includes those professionals for whom this measure was originally built.

MEMBER BRADLEY: Intended.

DR. BERG: And so as I indicated in our introduction, we built it specifically for the physical therapists, and then as was mentioned chiropractors were added. And then it became part of the PQRS Program after the original development of the measure.

MEMBER BRADLEY: Okay.

CO-CHAIR PARTRIDGE: I see a card up at the end of the table.

MEMBER AVERBECK: Yes, one other comment. As this measure concept matures to take a look at both pain and function, because it could be that either decrease in pain but may not be completely pain-free, and that balance that sometimes it's around improving your function and balancing that with the level of pain. Thank you.

CO-CHAIR PARTRIDGE: So it's been a long discussion, and as you predicted when you're the first measure up you've probably got a lot of questions that will be abbreviated as we work on some of the others, but frankly, I doubt it.

We were scheduled to take a break, we are going to take a break. I'm going to ask that in order to keep us more or less close to our agenda's time frame that we make it 10 minutes rather than 15, so back at 11:10, if I'm seeing the clock correctly, and same for our folks on the phone. Thank you for staying with us.

(Whereupon, the above-entitled matter went off the record at 11:00 a.m. and resumed at 11:14 a.m.)

CO-CHAIR STILLE: Okay. Welcome back. We are going to begin discussion of Measure 2967, the package of Home and Community Based Services Experience of Care Measures, developed by CMS.

Lead discussants will be Lee and Nicole, and I think Len, right? And Liz Mort also on the phone. And we'll have a three minute introduction by the measure developers. So, welcome, and take it away.

DR. LIDA: Good morning. Can everyone hear me? Okay. I'm Kerry Lida with CMS. And I'm in the Division for Community Systems Transformation and the Disabled and Elderly Health Programs Group. Guess not. Okay. Let's try it again.

And I'd like to briefly introduce our team who's been participating in this, and working on this. The experience of care area of work is about five years in existence at this time.

And so, I'd like to introduce Dr. Beth Jackson with Truven Health, and also Dr. Coretta

Mallery, with American Institutes of Research, AIR.

On the line we should have other colleagues from CMS joining us, including Jean Close, who's our deputy director for this division. And in the room also we have Allison Weaver, who works with this area of work with us.

And I'd like to briefly introduce some of the background information. And this will be very short, again, two minutes. Why we developed this area of work.

There is no cross disability beneficiary reported measures of service quality or experience with Medicaid HCBS, Home and Community Based Services.

And again, this developed over five years, beginning with NQE, the National Quality Enterprise, a grant. And most recently with the TEFT demonstration, Testing Experience and Functional Tools.

The development testing of these measures is directly related to a measurement gap

that was recently pointed out in the 2015 report by NQF. Can everyone hear me?

CO-CHAIR STILLE: You're a little soft.

DR. LIDA: Okay. And the quality report notes that there is a lack of systematic measurement of quality of HCBS across payers and delivery systems.

A brief overview of who the population is that we're serving, home and community based services. Persons receiving supportive Medicaid services so they can live at home and participate in community life and employment.

I know a number of you are very familiar with the recent shift in spending that we identified also within the background for this. We are over the 50 percent mark of spending for home and community based services, as opposed to the institutional services at this time.

Again, as it was noted, we submitted a package, including 19 experience of care measures, including seven scale measures, global

rating measures and recommendations, and individual items measures.

The importance of measures on Home and Community Based Services Experience of Care, CMS, and state Medicaid agencies both -- we require information on performance of Home and Community Based Services, providers and programs.

And the focus on the experience of care is very consistent with the recent rule in the end of 2014, that CBS rule, with the focus on person centered care, and person centered planning.

And the use of the measures that we are intending, states can voluntarily use these results for quality improvement and public reporting. We are using these in current demonstrations.

I noted the TEFT demonstration of which we have eight states participating at this time. This was part, and it was also continued from the NQE, where we had the field test data.

Round 2 for TEFT is going to be

starting very soon. And those eight states will be testing, will be using it again there. There's a broader interest of experience of care measures.

We have five states who have indicated they would like to use these at this time, including Pennsylvania, Massachusetts, Oregon, Nevada, Louisiana. We've had additional requests for information from MACPAC, AARP, researchers and MCOs.

We do plan a roll out of the data source for the experience of care areas of work, after we receive the CAHPS trademark, which is pending at this time.

And also, we have integrated the experience of care areas of work with a close project with the Office of the National Coordinator for an electronic long term services and supports plan that we're developing as part of the TEFT demonstration also.

So, thank you for your consideration of these areas of work. We sincerely appreciate

the opportunity to share this information with you today.

And I will turn to the next, I'll turn it back to our leads for the next steps, and what I anticipate will be a robust discussion that will continue. So, thank you.

CO-CHAIR STILLE: Okay. Thank you. So this is a package of 19 different measures, measured, I believe the unit of measurement is at the state level, what you've done you analysis as?

DR. JACKSON: At the state program level.

CO-CHAIR STILLE: That state program level.

DR. JACKSON: The state program level.

CO-CHAIR STILLE: Okay.

DR. JACKSON: Yes.

CO-CHAIR STILLE: Okay. Combining different providers that you've worked with?

DR. JACKSON: Yes. And there --

CO-CHAIR STILLE: Okay.

DR. JACKSON: There are multiple programs typically, HCBS programs per state.

CO-CHAIR STILLE: Okay. Okay. Great. Good. During the break we had talked about how to ideally try to get all of these discussed within an hour. With some preliminary comments that we've received, it's obvious that a few of the measures may require some separate discussion.

And what we've decided is that there may be one or more that actually require a separate voting. That if one requires separate voting, then we'll have to do a separate vote for all of the different parameters for that particular measure.

But then there are many others that may be able to be considered in one lump. So that's kind of how we're going to run things. I will turn it over to the lead discussants now to get into some of these things. And who would like to start?

CO-CHAIR PARTRIDGE: I will.

CO-CHAIR STILLE: Lee would like to start.

CO-CHAIR PARTRIDGE: I don't know that I'd like to, Chris, but --

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: I should just, in all fairness to our developers and to my colleagues, should point out here that I was a state Medicaid director for almost ten years.

And I -- in our particular jurisdiction it didn't have any HCBS programs, per se. Because we had an extremely broad benefit package. And most of the kinds of things that people were discussing at that point, that would be in a HCBS program, we were already approving as just standard care.

Since then the HCBS program I think has morphed into covering categories that most of us would consider marginally medical, a lot of physical adaptation, et cetera, all of which are good and critical to this program.

So, my first question, Chris actually

already asked, which is, we are not measuring at the state level, we are measuring at the state program level.

If you wanted to measure at the state level you would presumably take the average, let's say the state had ten programs, their average score. And that would be that state's score. And then we would look across the state. But we are talking at program level.

We have a 95 question survey? Am I reading that correctly?

DR. JACKSON: You are. You're correct. However, there are many skip patterns within the instrument.

CO-CHAIR PARTRIDGE: I know.

DR. JACKSON: So not all items are asked of the beneficiary, the program participant.

CO-CHAIR PARTRIDGE: Right. But for those of you who work with other CAHPS tools, and this one I know is up for endorsement, but maybe not there yet. Ninety-five questions overall is

a pretty big gulp, and can take a fair amount of time to administer.

And I think we'll get into those questions when we get to usability and feasibility. But, you know, right now when I looked at this initially, I thought, this is expensive. Particularly since we're still talking mail or telephone follow-up. And, Sam.

MEMBER BIERNER: I just, I'm still trying to understand the state program. Can you give me a feel for how many -- how big is the denominator? How many potential, are these each individual, you know, provider surveys, like a home base provider? Like, is there 100 in a state?

CO-CHAIR PARTRIDGE: No.

MEMBER BIERNER: How many are you talking?

CO-CHAIR PARTRIDGE: You're typically talking -- I'm out of date. I haven't been a director for a good many years. But I keep up. And I would guess that in a lot of states you're

talking of three, four.

It depends in part on the size of your state. VNA is often one of your contractors. You may, I don't know. Linda, you live in Massachusetts. You have some idea, I know.

MEMBER MELILLO: I'm not sure what level --

CO-CHAIR PARTRIDGE: Mic.

MEMBER MELILLO: -- it would be.

CO-CHAIR PARTRIDGE: Well, this --

CO-CHAIR STILLE: Linda, use your mic.

MEMBER MELILLO: Oh, I'm sorry. No. I'm not sure which groups within the state you would be referring to. Do you --

DR. JACKSON: May I answer that?

CO-CHAIR PARTRIDGE: Please.

DR. JACKSON: Yes. Frail elderly, working age disabled, for the most part. Okay. Physically disabled. People with intellectual disabilities --

CO-CHAIR PARTRIDGE: Okay.

DR. JACKSON: -- and developmental

disabilities. People with brain injuries. It encompasses those groups of individuals. And I think the working age disabled could include other, you know, other sub-groupings of people with physical and/or cognitive disabilities.

MEMBER MELILLO: Okay. But at the state program level.

DR. JACKSON: State program level. And they're -- some states may have four or five. There are some who, up to 11 of these separate programs.

MEMBER MELILLO: Okay. So, like the Massachusetts Brain Injury Association?

DR. JACKSON: It's brain injury waiver typically, the 1915(c) waiver programs.

MEMBER MELILLO: Yes.

DR. JACKSON: Yes.

CO-CHAIR PARTRIDGE: Okay. Great. Thank you.

DR. JACKSON: Although there are other authorities as well.

CO-CHAIR PARTRIDGE: We are going to

talk about these as a group. So, as we think about the different measures, we've got five sets, I think. The data source is the same. It's this survey tool.

Okay. I guess, at what point do we talk about which ones we might want to pull out? Maybe I'll just stop and ask my fellow discussants for comments that they might have. Most --

CO-CHAIR STILLE: Okay. What, yes. Why don't we have general comments from the lead discussants, and then the discussants can recommend any that they really feel should be pulled out for separate voting. And then we'll open it up to everyone. So, Len.

MEMBER PARISI: Thank you. So, I guess I still need some clarification on the population. Are these beneficiaries that are enrolled in a managed long term care program? Or, I'm still not clear on who these are.

DR. JACKSON: It is both. Both fee-for-service as well as Managed Long Term Services

and Supports. There are well over 300 of these waiver programs, non MLTSS fee-for-service in the country, nationwide.

There are also Managed Long Term Services and Supports programs that were part of the testing, and for which these measures would be applicable as well.

MEMBER PARISI: So, in a -- I'm already at the point where I'm trying to understand what we're really evaluating. So, at the end of the day the goal is to evaluate a Medicaid program?

DR. JACKSON: Yes.

MEMBER PARISI: For each state?

DR. JACKSON: Medicaid programs typically --

MEMBER PARISI: Not, okay.

DR. JACKSON: In a state where you have a unified Managed Long Term Services and Supports program it may be one program that you're evaluating.

However, the measures could be applied

and used across MCOs to compare the experience of beneficiaries in those different managed care organizations that are part of the Medicaid MLTSS program.

MEMBER PARISI: So, that's how I was looking at it. But it seems like it's more than that.

DR. JACKSON: It's both and.

MEMBER PARISI: Okay. And these are not certified home health agency patients? They're not on a short episode of care? These are long term?

DR. JACKSON: Yes. This is long term home and community based services to people who would typically otherwise be being served in an institution. So, it is long term services and supports.

MEMBER PARISI: So, there would be no overlap. What if they were on an episode of care in a certified agency while they were still on this program? Would they be excluded in any way?

DR. JACKSON: It's -- they may be in,

may be receiving those post-acute care services. But in general they're enrolled in this program. They met the eligibility criteria.

Say there is an episode, they go to the hospital, and they need to go to rehab. And then they would go back home.

MEMBER PARISI: Thank you.

CO-CHAIR PARTRIDGE: I'm now getting totally confused. I thought the entity we were evaluating was typically a subcontractor, an entity with whom the state can, Medicaid agency contracts to run an HCBS program for X, whether it's your DD population or your physically handicapped, or --

But if this is somebody who's enrolled in a Medicaid managed care program, like a dual, well, forget it's a dual, just Medicaid managed care program, the entity you would be measuring would be the MCO's performance, with respect to that segment of their population?

DR. JACKSON: I want to distinguish between managed care for the Medicaid population

in general, and Managed Long Term Services and Supports programs, which may include both long term services and supports, as well as healthcare.

And some states roll the two into one program. Some separate out the MLTSS. But we're talking about the MLTSS portion. If they're being served. If their healthcare is being provided as well as long term services and supports. So, we're parsing out the HCBS portion.

CO-CHAIR PARTRIDGE: But, you're then, if the -- okay. I just, I tend to go back to thinking, who's the provider in front of me? And what's the score for that provider? And --

DR. JACKSON: And this is a score for all homemakers that are serving that program, that HCBS, say waiver program. For people who are being served, a waiver that's serving --

CO-CHAIR PARTRIDGE: But those homemakers are being hired and paid by some entity.

DR. JACKSON: Typically, right. But they're typically --

CO-CHAIR PARTRIDGE: It could be the state itself. I understand that.

DR. JACKSON: It could be. It could be.

CO-CHAIR PARTRIDGE: But, okay.

CO-CHAIR STILLE: Okay. Len, did you have some other questions before we --

MEMBER PARISI: I guess, I guess -- and that's a really important point. Because, I mean, the goal of measurement is to have that be actionable so there can be some improvement.

And when it's -- in the case that you just described if you have multiple agencies that are providing various homemakers or staff to beneficiaries, it's hard to make that, those measures actionable to improve. Because you don't where to go with that.

So that was one of the things. As I was reading this it wasn't clear how you can actually do something with this information.

DR. JACKSON: The accountable entity?

(Simultaneous speaking.)

MEMBER SALIBA: This is Deb Saliba. Can I slip in for a second, to say that, you know, from California's perspective, I'm actually working with the state right now. And sort of trying to come up with a more unified approach to home and community based services.

The, I think you want to have some sense of how the state is doing and organizing, and setting up its programs for Home and community based services Based Services.

So you're right. It's often a hodgepodge of a lot of different programs. But from the consumer's perspective what matters is whether their needs are being met. Not so much which programs.

And if we're seeing variations that, you know, in some states that have a more -- let's just say is systematically organized, the systems tend to have better performance than in regions where there's less systematic

organization. Or vice versa, where there is sort of the more hodgepodge approach, might actually have higher levels of satisfaction.

I think that would be the way I would, you know, I think that would be very useful and helpful information from an HCBS, the Home and community based services perspective.

Though I think we -- I agree with you that it would also be nice to be able to look at it at the provider level. But, for example, the largest program in California is a program where the consumer gets to hire their own -- basically select their own person to do their in home care.

And so, you know, we're really not looking at a specific provider of care. I mean, the state directly pays those people. But it's the consumer that's selecting them and supervising them.

So, I think there's a lot of variability. But we really need some way of understanding what's going on in that space.

CO-CHAIR STILLE: Great. Deb, thanks.

That's an important thought. And please hold that thought. Because I think we can talk about that during the usability and use section.

I'd like to get Len, and then Nicole to kind of do their -- sort of finish their lead discussant stuff. And then we'll move forward. Anything else?

MEMBER PARISI: The other, I'm focused on the population. Because it wasn't clear who we're actually trying to evaluate in terms of the services that they receive. But it was clear, based on the discussion, who that is now. So, thank you.

CO-CHAIR STILLE: Okay. Nicole.

MEMBER FRIEDMAN: And I don't know if these will be more appropriate to get into further in the usability. But also thinking about the population.

And if we're evaluating homemakers, and caregivers, and different entities, how does this measure? If there's different levels of care that the state deems, you know, a certain

amount of hours a person gets, or based on what their need is, how does this measure look at that?

And especially around the component around unmet needs. Knowing that it might be that, you know, the person doesn't qualify for a certain amount of hours. Or there's variability in the amount of care in home.

DR. JACKSON: It does not control for that. What the measures do is tell you whether or not the person's needs are being met, at whatever level of service provision has been authorized for them.

CO-CHAIR STILLE: Nicole, any other points as lead discussant that you wanted to make? Questions?

MEMBER FRIEDMAN: I think mine are more for the usability.

CO-CHAIR STILLE: Okay. Great. So we'll do that then. Good. Okay. Thank you. Sorry.

MEMBER MORRISE: I want to point out

that my daughter was the recipient of a home and community based waiver. I'm very familiar with them. I actually worked in the managed care division of Utah Medicaid for awhile in the home and community based services area.

And while this does not apply to individuals under the age of 18, I have a number of friends who graduate into adult based waivers, based on disability services, traumatic brain injury, as you suggest.

Individuals in our state also are able to contract privately with providers. And then, based on who their managed care provider may be, they also go through agencies for services. It depends on how the specific waiver is written, and what services they are able to access under the waiver.

There is a real need, on an aggregate or composite, to determine if these services are efficacious. And so, I really liked this when looking through it.

I can see where a state could

potentially find out on an aggregate that needs were not being met. And then implement some kind of effort to mitigate that. But they wouldn't know, again, they won't know if they don't ask.

And these services are vital to being able to be in the community for so many of our disabled, but able to function individuals, with minimal support in the community.

And I'll tell you, these kids hate to still be living with their parents because they're the only ones who will take care of them. And these kinds of waivers get them out.

But unlike when my daughter was on the waiver, and I was managing the providers and making sure that the people coming into my home were doing what they were supposed to do, or then reporting them, the kids don't always have the same ability to follow-up unless they're asked. And I think it's really important.

And when I say kids, I'm sorry, I am getting seasoned. I'm talking about mostly young adults, 20s, 30s, who are able to be in the

community thanks to these waivers.

CO-CHAIR STILLE: Okay.

MEMBER FRIEDMAN: That's all.

CO-CHAIR STILLE: Great. Thanks. Sherrie, and then Nicole. And then we'll start to talk about evidence.

MEMBER KAPLAN: I'm still confused about the attribution, where -- what's being evaluated, what's the unit being evaluated. Because it sounded like, it's clearly the state's program that's being evaluated.

On the other hand, if there are, as suggested, and maybe I heard this wrong, some states have one, and some states have multiple. And so, the sort of nesting, you're going to confound states with programs that only have one program.

And for states with multiple programs, how much of the variation within -- where is the attribution going? Is it to the -- because some components are going to attributable to the state.

But, and they're, if you've got multiple programs it's more easy to evaluate. Is there a lot of between state, within state variation over programs? And how much of this belongs to things that the, you know, the state's program initiatives, et cetera.

Are you going to be able to distinguish the state variation from the program variation? And is there going to be a lot of reliability issues when we come to those, in estimating the within versus the between program variation within a state, across states?

DR. MALLERY: So, I think that's important to point out right at the beginning is that these will be administered moving forward by the states. So they might be administered differently within each state.

So, I don't think there is a plan to look at, to compare states at this point. So, it's almost like we have to look at within each state what is the different state's plan. Because they are very different. So, there isn't

a current plan to compare states.

CO-CHAIR STILLE: Okay.

DR. MALLERY: Does that help clarify?

MEMBER KAPLAN: Yes. I mean, I was trying to sort of get a sense of where, what your, ultimately the intent is. But if it's still -- you're not going to quite escape the issue of the confounding of state with program --

DR. MALLERY: Right.

MEMBER KAPLAN: -- for states that don't have very many of these things versus states that have multiple of these things. And then you can really look at, you know, is there variation across the state.

So, if you're doing it within states, that's even kind of a more perplexing problem, or potentially perplexing problem.

DR. JACKSON: I'd just like to add too, that this is going to -- this would be voluntary.

CO-CHAIR STILLE: So, Nicole and then Beth. And then I think we kind of need to move

on to the different parts.

MEMBER FRIEDMAN: Sorry. For the population that you would be excluding getting the survey, it says that it's folks that cannot pass cognitive screening. And at least in my experience a lot of the frail elderly who are getting in home service do have cognitive impairment.

Would you consider including caregivers or family members? Or how are you going to include this, you know, in my opinion big population that we want to have a -- to be heard?

DR. JACKSON: For the testing, because we were going for a CAHPS trademark, and they historically have not allowed proxies, for the most part we excluded them.

It became clear as we were progressing in the field tests that proxies were popping up. And so, we did allow them from a certain point onward.

What we are doing in the

demonstration, in the TEFT demonstration for Round 2 of data collection is allowing the TEFT state grantees to experiment with that. So moving forward there's an acknowledgment that proxies may be necessary.

DR. MALLERY: Can I just clarify that? The proxy data are not in the results that are presented in this packet. So we would have had about 1,000 more responses if we would have included. But because it wasn't administered to proxies consistently throughout --

CO-CHAIR STILLE: Okay.

DR. MALLERY: -- we excluded those from these analyses that are reported here.

CO-CHAIR STILLE: Okay. So your data in which you're applying for is without proxy?

DR. MALLERY: That's right.

CO-CHAIR STILLE: But maybe that might change. And then what does TEFT mean? I'm sorry.

DR. JACKSON: Testing Experience and Functional Tools.

CO-CHAIR STILLE: Okay. And, Beth, you've been waiting patiently.

MEMBER AVERBECK: I think, Lee, you had mentioned earlier on whether or not we pull any of the questions out separately. And I don't if we're getting to that, or if you want any comments now. So, I should wait?

CO-CHAIR STILLE: Just wait 30 seconds.

MEMBER AVERBECK: Okay, great.

CO-CHAIR STILLE: Okay, great. We'll start to discuss the evidence now. And I'd first ask if any of the lead discussants feel strongly that we need to pull one or more particular measures out of this package.

CO-CHAIR PARTRIDGE: I think I would, I'd be more comfortable if we pulled the physical safety question out as a separate one. Because it seems to me it gets close to the never events issues.

CO-CHAIR STILLE: Okay. And that, is that Number 19, I guess, in the list they have?

Okay. Others?

CO-CHAIR PARTRIDGE: The other possibility I would maybe note, and I'm not sure. When you look at the performance gaps, as I read the data it wasn't significant on all but the questions relating to unmet needs. And then I said the physical safety one.

There are the five measures relating to unmet needs. And I don't know whether that's appropriate to pull them out separately, or just raise it when we talk about the whole package. Sarah?

CO-CHAIR STILLE: Okay. So --

CO-CHAIR PARTRIDGE: I don't want to see five separate ones.

CO-CHAIR STILLE: Right. So potentially the unmet needs is a package.

MEMBER PARISI: So, I hate to ask the question. But I'll bring it up. So when Sarah had talked about the concept of approving, or I think maybe you did Chris, approving the tool versus approving the criteria. So, we're looking

at this as 19 different measures, and not as a tool, right?

CO-CHAIR STILLE: Right.

MEMBER PARISI: And that's how it's going to be loaded? Okay.

MEMBER BEVANS: That almost got at my question I think. But, just so I understand as a point of process. If we were to approve this measure set, would that be like NQF saying that use of this measure must, is recommended to be used as a full measure set? Or does that mean that it could be used piecemeal?

MS. SAMPSEL: Yes. So this is similar to, in Phase 1 with the CAHPS measures. And what we would see is different programs report different measures. So, we would be, you know, potentially endorsing 19 measures.

But that's what we're trying to get at now. Are there any you would like to pull out that make you very uncomfortable on making an endorsement decision on? Or do you feel that there needs to be a discussion?

What happened in pediatric project, I believe, Suzanne. And maybe, Suzanne, why don't you talk about how it was handled in pediatrics.

MS. THEBERGE: Sure. We had a measure. It was similar to this. And we ended up, the committee requested that the developer split it into 19 separate, or sorry, it wasn't 19. It was ten in that case. Ten separate measures.

And the committee voted on each one separately. Eight of them were endorsed. And in between the -- before the public comment period we went on the back end and split them out. So they became ten separate measures, with ten separate numbers. And the sub names of measures were the new title of the measure, et cetera.

MS. SAMPSEL: But, and back to your question as well, Katherine, is, I mean, I think states in turn, and I think you guys should answer this.

States in turn could say, we're only in our -- on our state dashboard for HCBS cert

programs we're only going to report the global ratings measures. Even though all are endorsed, states may choose to only report some of them.

CO-CHAIR STILLE: Okay.

DR. JACKSON: I think that's correct.

CO-CHAIR STILLE: Others that feel strongly about separating any particular -- Len, you have your hand up.

MEMBER PARISI: Well, question one and two, when you say staff, I don't even know who that's looking at. Is it the homemaker? Is it the case manager? Is it -- it's very unclear. So, I'm not sure that that would be something that you could even have a discussion about it, in terms of voting.

CO-CHAIR STILLE: So, yes. That might be more for the validity discussion.

CO-CHAIR PARTRIDGE: Well, I'm going to try to help you out, Len. I don't think you would be directly measuring the individual homemaker. I think you would be --

In the, and I'm talking really very

much from a common sense, I used to run this program kind of perspective, and not from the measurement set issue perhaps.

If I had, as you do in Utah, and I guess a couple of other states, a family in which you can choose your provider yourself from an approved list, supplied by the state.

I would say in that case the entity you're measuring is the state agency itself. How well does it select the people with whom it contracts, monitor them, and so on?

They're being -- in my state, for example, it was run by our Department of Social Services. So that would be, you would be evaluating the state.

But if you turned around and contracted with say the ABC Home Health Agency to contract with that group of homemakers, and supply them, and pay them, and monitor them, and vet them, and so on. You'd be evaluating the ABC Home Health Agency. Am I right? I mean, CMS, am I --

DR. JACKSON: What I'd like to do is just direct us to the actual items, or questions in the survey that --

CO-CHAIR PARTRIDGE: They do get there, right.

DR. JACKSON: -- identify who that worker is. And it can be a personal assistant, a behavioral health staff, or a homemaker for this particular scale. So it is identifying in general an individual.

CO-CHAIR STILLE: But it's not reported like that in the data? Or just aggregated at least one level above that, right? Okay.

DR. JACKSON: It's at the program level --

CO-CHAIR STILLE: Right.

DR. JACKSON: -- that it's reported.

CO-CHAIR STILLE: Right.

DR. JACKSON: Yes.

CO-CHAIR STILLE: Okay. Sherrie, you had a point?

MEMBER KAPLAN: Just one point about if states can cherry pick off from these things? The single item measures are notoriously unreliable.

So it does change the conversation about whether we want to include those as a group discussion and, you know, because the reliability issues do change when you have single item measures that could be potentially lifted out of this.

If you're going to, you know, attribute to the program, and it's got all this kind of variance that's attributable to the program from the kinds of people that are contracted with the providers of the program, yada, yada, single item measures really have that as a separate issue.

CO-CHAIR STILLE: And the single item measures would be 11 through 19, I guess, right? Yes. Okay. Nicole, you have a thing up?

MEMBER FRIEDMAN: I just had a question to clarify. So for the scale measure,

choosing services that matter to you. Aren't these services that are --

As mentioned, there's not a lot of choice. That there's kind of a bundled element of services that Medicaid will pay for. And so, I just don't see how, if there's, how the recipient, or the customer would be able to have choice of the services.

DR. JACKSON: There is typically a package of services per program that the individual can choose from. So for example, personal care. An individual may have a personal assistant come into the home. Or they may go to an adult daycare to receive those services. So it's that kind of choice we're talking about here.

CO-CHAIR STILLE: And then, Jennifer, did you have a question?

MEMBER BRIGHT: Yes. Just, first a point of clarification. We're still talking about how to approach these, as opposed to the evidence part, right?

CO-CHAIR STILLE: That's right.

MEMBER BRIGHT: I wanted to make sure I ask a germane question --

CO-CHAIR STILLE: Yes. That's right.

MEMBER BRIGHT: -- before I do anything. So, this is a question to the developer. Can you differentiate for me why we're talking about both a global ratings set of measures, and a recommendations measures? Because to me they kind of get at the same thing, but just at a different approach.

And given that one of the immediate concerns is the scope of the scale of this, why are you trying to lump so many things under one umbrella, is one question.

And the second is to reiterate the previous comment about individual items. I would say we should be talking about these as kind of buckets. I mean, there's, the scaled measures are really about specific measures of quality of the services, and types of services that are being provided.

The global ratings and recommendations measures are really patient satisfaction. How do you feel about what you're getting? And the unmet needs is trying to assess where there are gaps.

So, if we're trying to figure out a way to break this up, I would posit that we should be thinking about them in those buckets, as opposed to pulling them all apart into 19 different painful discussions.

CO-CHAIR STILLE: And so we have five buckets with Number 19 in its own bucket. And, you know, that's interesting. So, considering these as five groups of measures, I think would answer everyone's concerns, from what I've gotten so far.

MS. REED: Chris, can I just follow-up?

CO-CHAIR STILLE: Yes.

MS. REED: Because this, I didn't want to be misleading. When, like the CAHPS measure, I was involved in the initial CAHPS measure, way,

way, way long ago.

When we did that you put, you sort of general overall ratings and recommendations that's what we call behavior intentions. The latter being behavioral intentions. And the global rating items are used as validation items for these subscales.

So, the scales for, you know, the staff are reliable and helpful, and the staff listen to me, and everything, those sort of overall ratings are usually used for validation internal to the instrument construct validation.

So, we want to make sure we're measuring where we -- so, you wouldn't sort of -- that's why I freaked out. You wouldn't --

CO-CHAIR STILLE: Right.

MS. REED: -- separate those from the interior of the subscale structure that tells you where you need to focus the improvement. Because you want to use those measures to find the drivers of the overall ratings. For example.

CO-CHAIR STILLE: Right. So don't use

one without the other is kind of what you're saying. Yes.

MS. REED: I was just saying it wouldn't be really, I would feel uncomfortable breaking those out as a separate set when they're usually used in the context, especially HCAHPS and CG CAHPS, and so on.

They still put those, you wouldn't ever separate those out.

CO-CHAIR STILLE: Right. Linda. While we're catching our breath and figuring out what to do next.

MEMBER MELILLO: So, I'm going to muddy the water a little bit. Community services frequently offer employment services as well. I didn't see that anywhere.

DR. JACKSON: You didn't. It is in the instrument. There is what we're considering now a supplement to the instrument, an employment supplement.

And the reason that it really was not tested, because so few people who were surveyed

answered the question in a way that said, yes, I'm either interested in employment, or I am working. And you had to -- that was the trigger for those -- for that battery of questions.

And there were so few who said either, yes, I'm interested in working, not currently working, or I am currently working, to administer that supplement. So it is a supplement. But we did not test it. And I'll defer to my statistician here.

DR. MALLERY: So, those items will still be part of the instrument that we're not technically talking about here, but it will be part of the survey. So states will be able to still use those items. But we're not putting those up for endorsement, because we just had too few respondents to be able to sort of report reliability and validity --

MEMBER MELILLO: So, how many additional questions is that?

DR. MALLERY: I can tell you in a -- give me a minute. But it -- yeah.

DR. JACKSON: Again there are a lot of skip, there are skip patterns. So.

CO-CHAIR STILLE: Okay. While you're looking, what we've decided to do -- this has been a great discussion, by the way, so, thank you. We're going to take any more general questions, including about, you know, splitting up discussion. We'll take public comment. And then we're going to break for lunch and discuss how we're going to vote on these while we're eating lunch.

We're going to take a 20 second time out, is really what we're going to do. So a little more than 20 seconds. 20 minutes? I don't know.

MEMBER THOMAS: I haven't said anything. Oh, I'm sorry.

CO-CHAIR STILLE: Okay. Jennifer, I think you were first.

MEMBER BRIGHT: I just wanted to go back to the first question I posed to the developers, which is, can you help us understand

the difference in the global ratings measures and the recommendations measures?

Because, again, I think they're trying to get at the same thing. And what's the value of having both? Because perhaps there should be some consideration of consolidating. I don't know.

DR. MALLERY: So, I would argue to keep both. I think they are used somewhat differently. And the global ratings items, again, are often used to correlate. I think that -- I mean, I guess they both could be used for that, for those purposes. But I do think they could be used slightly differently. But I do understand your point that they might be getting at the same concept.

CO-CHAIR STILLE: Okay. Becky, and then Nicole.

MEMBER BRADLEY: I just had a question about the intended use. Do you intend this data to be publicly reported in any way? Because, unlike some other providers, services, patients,

or communities, I mean, their only option would be to move to another state where there's a better program.

So, how do you intend to use the data from this? And I guess that gets to intended use. But I apologize for that.

DR. MALLERY: Yes, thank you for your question. And at this time we're just looking at states who are voluntarily using it. And so it would be their option for the TEFT states that we have. We have eight TEFT states who will be using this. TEFT again was Testing Experience and Functional Tools. It is their option if they would like to publicly report it.

The initial round 1 of the data, we did not publicly report it. It was available in individualized reports to the states. And then we shared information on -- aggregate information, but the states did not want that information reported. They wanted to keep it for their own purposes. And we respect that. But again, that was just for the TEFT demonstration.

We're doing the same thing for Round 2. No plans have been made past that, beyond that. You know, we have five states who have expressed interest in using this. That would be their option. I don't know if Beth or Coretta have information to add to that.

CO-CHAIR STILLE: Okay. Peter.

MEMBER THOMAS: Just really quickly, as a general comment. I would think this -- first off, this is an area that is just tremendously important for the disability community, the fact that the 50 percentile has been breached and we're now in a situation where we could provide care under Medicaid through home and community-based services more often than through institution-based services, especially for some of these populations, like brain injury and working age adults with disabilities, and the like.

So, this is a critical area that I'm really excited that we're kind of considering this. And I can most definitely see the states

taking a real stock of their different programs, their different waiver programs.

And there's so much variation in the states on how these programs are run that it can help the states determine whether the program, and the goals that they need to achieve with a particular program, are actually current, whether the end users feel that they're empowered to hire their own personal assistant. Or whether they've gotten too many restrictions in the program that it's not working the way they envisioned it.

And that's really valuable information that I think states can use to improve all of these different types of programs. So, tremendous variation across the states. But I can most definitely see these measures being very useful, not only for the states, but for the end users.

CO-CHAIR STILLE: So, good. So, where I would like us to be when we finish talking in the next couple of minutes is ready to come back from lunch to talk about evidence, validity,

reliability, usability, use, and that kind of thing, in whatever buckets we end up talking about.

So, other general questions before we open it up to public comment? So, should we? Okay. People on the phone?

MEMBER CELLA: None here, thanks.

CO-CHAIR STILLE: Okay.

MS. THEBERGE: So, public comment. Operator, can you open the lines for public comment? And if anyone in the room would like to make a comment, there's a microphone.

OPERATOR: Yes, ma'am. At this time if you'd like to make a public comment, please press star then the number 1.

MS. THEBERGE: Comments can also be submitted via the chat function on the webinar.

OPERATOR: And at this time there are no public comments from the phoneline.

MS. THEBERGE: Okay.

CO-CHAIR STILLE: Okay. So we shall break for lunch. And let's see, 30 minutes, 45

minutes? Okay. So if you could make it a relatively brief lunch. We have 27 minutes until 12:30 p.m.

(Whereupon, the above-entitled matter went off the record at 12:04 p.m. and resumed at 12:34 p.m.)

CO-CHAIR STILLE: All right, great. So, over lunch I was made aware of the possibility of deferring a vote on this package of measures. And it would be good to have a discussion of that. But we need someone to move that first.

So, Jennifer, would you like to kind of talk about kind of your thinking behind that? And then we can open it up to some discussion.

MEMBER BRIGHT: Yeah, so I'm guilty of raising this question. I just, as a new committee member, was asking procedurally, was there an opportunity to discuss the pros and cons of possibly deferring the consideration of this measure?

I think there are sufficient questions

about scale and about lack of data, being that this has just been field tested. And I think this is not a question of value. This is not a question of the value of doing this, of the considerable amount of work that's been put into this. But I feel that it's important to move to consider deferment because I think that this needs to be tested and teased out a little bit more.

And I do feel like the demonstration project that you mentioned in your introduction has a lot of value to answering some of the questions that I think have come up already and are likely to come up in further discussions.

So, for that reason, I'd like to suggest to the committee that we consider the pros and cons of deferring this at this time.

CO-CHAIR STILLE: So let's discuss this a little bit. I think I saw a couple of tents going up at some point. No? Yes? Lisa?

MEMBER SUTER: So, can I just clarify, what are the -- what's the downstream effect of

deferment? And what are we asking from the developer that would change our mind in the future?

MS. SAMPSEL: So, basically, I mean, what we would do now that there's a motion to defer, there would be kind of a discussion. And it can be some back and forth continuing with the developer on, you know, what your global reasons are for suggesting the deferment and seeing if the developer has some information we're just not seeing here.

And then there would be a formal vote. And it would be a hand vote of who supports deferment, who doesn't support deferment. And should that deferment move forward, then we would still want you to go through each of the criteria and talk about what additional information you would want to see that would make you more comfortable with a vote, so that the developers would know and be able to come up with a plan that would then come back on how they would address those issues.

CO-CHAIR PARTRIDGE: I'm sorry Sarah. Would we vote on the motion to consider to deferment? Is that it? Or do have to decide to defer first, and then, as we go through the criteria, we're giving them suggestions? Or do you assume the general discussion will surface the issues?

MS. SAMPSEL: I mean, I assumed that. How have they done it in other groups?

MS. MUNTHALI: Yeah, so you would vote on the motion to defer first, to even consider that. This measure could go forward if it passes all of the major criterion, but you're not sure. It sounds like you're saying you like the idea of this measure but there may be some weaknesses that the developer may be able to address within a time period.

And so when you're thinking about deferment, and when we're thinking about it as an option for committees, we would be working with the developer to come back with that additional information that Sarah mentioned within a

specified period, no more than a year.

So that's something to consider for the developers as well.

MS. SAMPSEL: And what I would just add there, too, I mean, the other option, of course, would be to continue and we vote through. And let's say some of these go down, they don't pass. Remember there's almost that cure period where the developers could come back and say, we want you to reconsider this, reconsider our request.

But there's only about two months for that. So it's during the writing period, it's during the public comment period that they could come back to us with a formal reconsideration and you would re-vote in about two months.

CO-CHAIR STILLE: Peter?

MEMBER THOMAS: I guess I would speak in opposition of the motion because I don't think we've really had the discussion to the point where I could say one way or the other that I think you need to develop more things. I've read

the materials. This is happening. These programs are in effect. Half the Medicaid population is undergoing some form of home and community-based services and there's very few measures out there.

So it's really important. And I don't know if a year delay is appropriate in light of the discussion we've had thus far. Maybe if we have more discussion I can get my arms around that. But I just don't know enough right now to say that. I'd prefer to move on with the discussion of the measures.

CO-CHAIR STILLE: Other thoughts?

CO-CHAIR PARTRIDGE: Sherrie's tent is up.

CO-CHAIR STILLE: Oh, Sherrie, sorry.

MEMBER KAPLAN: So can you go over, Sarah, the algorithm for our decisionmaking again? If we say deferral for a year, then what happens? If we move on and discuss these and decide, for example, to approve some and not others and do whatever it is we do after the

discussion, what happens? So that we can kind of clarify that. Because I think in one case, there's a short term and then other cases a more long term plan, it sounds like.

MS. SAMPSEL: So there would be a vote on the motion to defer. If it's decided that you all want to defer this measure, then we would still ask you to go through the criteria and provide recommendations to the developer on what you would want to see to make you more comfortable that this measure is ready for prime time, or measures are ready for prime time.

In that case, then the developers will be followed up by staff within 24 hours, 48 hours, when we get out of this meeting. And the developers then have two weeks to provide us with their plan to address those issues and a timeline for when that would happen. As Elisa mentioned, we'd want that to happen within a year.

Should you not defer that, you know, as Peter suggested, what we're going to do is we're going to break these into five groupings

and talk about each grouping and have five separate votes.

And let's say in the event, you know, say physical safety, Number 19, it's voted down. And this could be the case for any of these groupings. If any of them go down, as in are not passed at any of the must-pass criteria, we would still, as we've done before, we continue the discussions. But we don't continue voting for each of those additional criteria. Again, to give the developers more information.

In that case, where either the consensus is not reached or the measures are not passed at this meeting, they still go out for public comment. And the developer still has a request or is still able to bring more information back, but it's on a much shorter time period. And you would re-vote at the post comment call. Is that clear?

CO-CHAIR STILLE: Linda?

MEMBER MELILLO: So I was just -- I feel like that it's a little premature to defer

without having had more discussion about some of the actual items. Is there a way to continue to review the items and then decide whether or not we -- you know, that that could be an option at the end, is to defer?

MS. SAMPSEL: I think we can do that. I mean, as long as it's before the overall suitability vote, the very last vote.

MEMBER MELILLO: So that's my thought. I don't know how the rest of the panel feels.

CO-CHAIR STILLE: So should we maybe have a structured discussion, relatively briefly, on sort of the five different chunks? You know, the evidence things and the usability and use things. And then maybe have a vote to defer based on how that's going? Should we give ourselves like a time limit perhaps? Okay.

So we're supposed to be starting on the other measures, like, in two more minutes. Okay. Well, this is important enough. Let's give ourselves 15 minutes. Does that sound okay? Is that too long? Okay. Len?

MEMBER PARISI: So, clarification, yet again. What's the implication of looking at this document in buckets? And if one component is not acceptable, what's the implication of that?

CO-CHAIR STILLE: For the others? Yeah, that's a good question.

MS. SAMPSEL: So, I mean, I think what we've heard is that, you know, there seems to be a level of discomfort in voting on 19 measures separately and doing 19 separate votes. And so our suggestion is, we'll talk about the scale measures and go through the voting on the scale measures. And then do it on global rating, do it on recommendations, do it on unmet needs.

In the event any of these buckets or individual measures are not passed, it's the same implications of that, you know, there could be something that comes up during public comment. There would be a reconsideration of that vote post-meeting.

CO-CHAIR STILLE: Okay. Great. Well, let's do this: let's start to kind of go through

one-by-one with an eye toward, okay, what's needed, what's not needed for these generally. Because obviously they're all related because they're part of the same instrument.

Should we maybe get up the evidence thing and start going through there? And I'll rely on the primary discussants to put in a point or two to kind of jump start the discussion. Nicole?

MEMBER FRIEDMAN: I guess one of my thoughts of this is some of the questions on the scale measures are pretty similar to Press Ganey in like, staff reliable and helpful, listen and communicate. And I'm just thinking about how many surveys people potentially get. And would this be duplication of other surveys that they're getting, whether or not their home and community-based service is separate from their primary care health system. And is that a concern?

DR. JACKSON: Typically, an HCBS program would only field a survey like this, at most, once a year. Certainly, individuals who

receive their services, you know, from primary care or a specialist, go to a hospital, of course, they're going to be receiving surveys as well.

But these will be different in that, for the most part, they will be administered face-to-face or over the phone. So it's a different type of survey than what most of us and most of them will receive from our healthcare providers.

I think the states who would be responsible for paying for, for the most part, and choosing a survey vendor, and fielding these surveys with their beneficiaries would be very careful about participant burden and would not do it very frequently.

Plus it would be a cross-section, it would be a sample. It would not be everyone, in most instances.

CO-CHAIR STILLE: So, again, this is sort of evidence that this is needed and that kind of thing. So, I mean, your question was

great. I think it was probably more about feasibility. Katherine?

MEMBER BEVANS: A question about the data presented for evidence for performance gap. I just want to make sure I understand the means that are presented on Page 6 of the report, which are things like staff are reliable and helpful, the mean is 93.23. Could you tell us how that was calculated? I understand that the measure itself is a composite. And then there must be a cut point and a percentage calculated?

DR. MALLERY: So each of the measures, or each of the items in the underlying measure, is on a "never, sometimes, usually, always" scale. And then we transform those to a 0 to 100 scale, primarily, so they're easier to understand.

MEMBER BEVANS: Is it a T-score?

DR. MALLERY: Between the means and the -- that's right.

MEMBER BEVANS: No? Okay. No.

DR. MALLERY: Oh, no, I'm sorry. So

you're saying is this a Z-score? Oh, right, no. It's just, you know, take all the items, they're "never, sometimes, usually, always." So, one, two, three, four. Create a mean and then transform that to 0 to 100.

MEMBER KAPLAN: I think I can help out here.

DR. MALLERY: Okay.

MEMBER KAPLAN: What's usually done in that kind of transformation is you take the item mean, or the average item mean, and you subtract the theoretical minimum from that. And then you divide that by the theoretical maximum of the scale score minus the theoretical minimum and multiply it by 100.

So it's not a transformation in the sense that you're thinking.

CO-CHAIR STILLE: We're finding the table that you're talking about. I know exactly -- there you go.

MEMBER BEVANS: Understanding that a little bit more, it might be worth noting that

the scores are quite high on the possible range of 100, with obvious exceptions, which suggests there may be not the strongest evidence for room for improvement. Although it's kind of surprising given the content of the items.

And so I'm wondering whether you have any thoughts about sampling, who actually responded to this, how the measures were administered, that may be explaining those extraordinarily high scores on some measures that we know there are certainly more variability in prior literature.

DR. MALLERY: Yeah, I agree, the scores are quite high. In terms of sampling, you know, I mean, it was a random sample of participants within the programs. Now, in terms of people that responded, there were those cognitive screening items. So that's a consideration. So, only respondents who screened into it past this sort of cognitive screening piece. And that was just -- they asked whether or not they received services. So that might be

one consideration.

In terms of -- I think we already said that all surveys were administered in person or via phone. So I think the administration was appropriate for the population. So that would sort of mean, you know, we made every accommodation to the participants in that sense.

CO-CHAIR STILLE: Sherrie and then Lee.

MEMBER KAPLAN: This is a follow-up to Katherine's question, I think. If you look at the interquartile ranges, you get fairly narrow -- you've got ceiling effect problems with some of the measures. And the standard deviations are pretty, you know, pretty tight.

So then the question becomes, is there going to be enough -- you know, an effect size, for example, of .3, which means you multiplied the standard deviation times .3 and you get 1. Well, is one unit difference a meaningful difference?

And so that's going to happen to these

kind of ceiling measures. Now, since this is voluntary, your data are voluntary, it's sort of like what happened the last time. You get high performers participating in these voluntary kind of exercises.

On the other hand, when you look down at some of the unmet needs things, you've got a huge standard deviation. And therefore, you know, you've got kind of the opposite problem. Some of these things you might even have floor effect problems. Especially, you know, for some of the measures.

So when we're kind of trying to wrap our heads around -- I think this is the issue that Peter brought up. When we're trying to wrap around, the collective here, there's a lot of differences here that we're going to have to struggle with. And how to help us, how you can help us understand the potential utility, that gets to the use, but it also gets to the variability and understanding the attributions of the program versus other kinds of things that may

be causing all of this wide variation in some of these measures, but not very much and may not even be meaningful in some of the others.

CO-CHAIR STILLE: Okay. So Lee and then Linda.

CO-CHAIR PARTRIDGE: If you step back away from the detail, and overall, as I've worked with these measures over the past two weeks, you know, there's no question that Peter is right. We would very much like to have solid quality measures in the HCBS realm.

It used to be as the states struggled with them -- and I'm sure some of these are still out there -- what you did was what I would consider sort of dumb stuff. You pulled a sample of the care plans and then you went and checked and found out whether or not Mrs. X got three hours or two-and-a-half hours of personal care each day. Well, that's silly. You need something more global.

In a discussion with a bunch of HCBS providers a few years back, unanimously they

said, you know, what we really care about is whether people are doing well. Doesn't really matter about all this detail in between. What happened? Are they still functioning okay? If they're not functioning okay, why aren't they functioning okay?

At that point, you start to need the detail underneath to say are they not doing well because the agency is letting them down and are not supplying, meeting their needs.

So, you know, everything in me wants to say, I want to bless this measure. Nevertheless, I kept looking at some of our criteria. The performance gap really isn't apparent, except with respect to the unmet needs group where you do have a lot of indication that people are feeling that their needs are not being adequately met. Some of that may be under the control of the program. Some of it may be under the control of the state budget office.

But I am really terribly, terribly conflicted on taking our standard algorithm and

working through this measurement set, whether we do it in chunks or whether we do it one-by-one.

CO-CHAIR STILLE: Just with the differences in data ranges, data analyses, data requirements --

CO-CHAIR PARTRIDGE: Right. And you know, if you look further in the testing realm -- and I'm jumping ahead a little bit. I was stunned by your comment this morning that there were 2,300 respondents in this material that you gave us. If you had counted the proxy data, you would have had another 1,000. Well, what did that look like? So.

CO-CHAIR STILLE: Okay. And then Linda. Then I think we should talk about some of the other realms just to get to the end of this 15-minute block so we can decide on whether to defer or not.

MEMBER MELILLO: When I worked for the behavioral health organization in Vermont, they had a community service program. And one of the challenges was that any time you have a staff

person who is working with a developmentally disabled young adult, older child, they are heavily influenced by that person. And if they're the person administering this set of measures, I don't know that you're going to get real values that way.

So I also wanted to see that other 1,000 data points to see if you're going to be getting different types of responses that way. I think the administration of the survey, and really narrowing that down, would be really important because there's just way too many variables.

DR. MALLERY: Can we clarify that the survey is administered in-person but not by the providers. It's administered by trained survey vendors.

MEMBER MELILLO: Okay. Thank you.

DR. MALLERY: So no case managers.

MEMBER MELILLO: Okay. And family members, would they be able to answer as well?

DR. MALLERY: Well, so, that's the

difference, is that initially, no, they weren't able to answer. So they would never be administering the survey. It would be a trained survey vendor.

MEMBER MELILLO: Right.

DR. MALLERY: But that was sort of the point we were getting to, is that part of the way through we did allow family members to respond, not on behalf, or along with or to assist with the survey process.

MEMBER MELILLO: Okay. And that data is not in this set?

DR. MALLERY: It's not in here, right, because it wasn't consistent throughout

MEMBER MELILLO: Yeah, I wouldn't want it to be.

DR. MALLERY: So we couldn't reliably say, you know, these are the differences between proxies and non-proxies because it wasn't -- that wasn't how it was designed, the field test.

CO-CHAIR STILLE: Okay. Any other last important comments on evidence?

MEMBER THOMAS: Only that this is a patient-reported outcome measure. And so, in terms of evidence, you're asking the very people who are receiving the services what they think about the various components of what they're getting. I'm kind of sensing that we're maybe making this too complicated. This strikes me as being a real step forward from where we are now. And I guess I'm kind of left with that as a bottom line.

CO-CHAIR STILLE: Nicole?

MEMBER FRIEDMAN: Sometimes, in my experience, too, working with patients and asking them about their service, especially around Medicaid and help in the home, is a fear of losing the service. And so I'm wondering, is there scripting, when you're delivering this, that this is not about whether or not we're going to take away hours or et cetera, but this is a care optimizing tool.

DR. JACKSON: That's part of the consenting process of the survey.

CO-CHAIR STILLE: Okay. So just in the spirit of dividing up chunks of measures versus not, Sarah was raising the point that we haven't heard anything in the discussion about -- at least on the evidence part, just thinking about evidence -- that there are big differences between chunks of questions.

So we were thinking we should vote on the evidence kind of by itself and then think about the next few things. Does that seem okay to people? Beth?

MEMBER AVERBECK: Specific to evidence, I guess that's my question around the 19 and the physical harm. I don't know if you want me to raise that now or if you want me to wait.

CO-CHAIR STILLE: Sure.

MEMBER AVERBECK: Because I think one of the -- you know, obviously, it's considered a never event. This is a patient-reported, so it's not necessarily a verified result. And so I think that may play into, you know, my

consideration on whether it's just reported and verified or just reported, especially if we're going to start looking at comparative data or public reporting.

CO-CHAIR STILLE: Over to Desi for a vote on evidence, then. And then we'll continue the discussions.

MS. QUINNONEZ: We are going to open the vote. Voting is now open for evidence, importance to measure and report. The rationale supports the relationship of health outcome, or PRO, to at least -- what are we voting on? Sorry.

Oh, thank you. I can get you the screen.

The rationale supports the relationship of health outcome, or PRO, to at least one healthcare structure process, intervention, or service. Option 1 is yes, Option 2 is no. Voting is now open.

MS. THEBERGE: We're down to 19 Committee members now.

MS. QUINNONEZ: Thank you.

CO-CHAIR STILLE: It doesn't seem open.

MS. QUINNONEZ: It's calculating. We're looking for one more vote.

MS. THEBERGE: Deb, are you still there? Operator, is Deb Saliba still on the line?

OPERATOR: I do not see her joined.

MS. THEBERGE: Okay. So I think we're good.

MS. QUINNONEZ: So we're at 18. All votes are in. So voting is now closed for Measure 2967 for evidence. We have 94 percent -- that voted no. That's 94 yes and 6 percent no. This measure passes.

CO-CHAIR STILLE: Are there other evidence sub-questions? Okay. Performance gap. Okay. Any other questions about performance gap related to this table? I think we've talked about it fairly extensively. Becky and Sherrie?

MEMBER BRADLEY: I just had one

question. Or comment I guess. In terms of now that you're allowing proxies, I think there might -- it's been my experience in working with patients and families that sometimes the caregiver's perception is very different from the patient's perception, in terms of adequacy of service.

And so now that you've added those kind of as one of the acceptable responses, I think there should be more research into, is there a difference between the patient's and caregiver's responses to adequacy of service? Because if you're just basing this on the patient, there may be a big gap there.

CO-CHAIR STILLE: And Sherrie?

MEMBER KAPLAN: Okay. So this is already where I'm hitting the wall on, is this a bundle or a bunch of individuals? Yes and no. I mean, if you stare at this, you know, what are we supposed to do? Because I can't answer the question.

CO-CHAIR STILLE: Right. Answer

Number 5.

MS. SAMPSEL: I mean, at this point we can separate. You can go ahead and say, okay, we're now going to do one vote on the scale measures, if that's what the suggestion is. Then do one vote on unmet needs, one vote on, you know, through each five.

CO-CHAIR STILLE: And I think we were kind of moving toward a consensus of that. So should do that, then? The five buckets being -- I don't have them up on my screen.

CO-CHAIR PARTRIDGE: Scale, global, recommendations, unmet needs, physical safety.

CO-CHAIR STILLE: Okay. Scale measures, global measures, recommend measures, unmet needs measures, and physical safety. So then we'll have five quick votes on 1(b), right? Okay. So, good. So this will be on the scale measures.

Is voting open yet?

CO-CHAIR PARTRIDGE: Desi, is voting open?

CO-CHAIR STILLE: I think it's not open yet. Yes, sorry. The splitting causes technical difficulties but we just want to try and get this right.

MS. QUINNONEZ: Okay. We will open up the vote for importance to measure and report for performance gaps. Performance gap data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups or disparities in care. And this will be for scale measures. And that's a part of 2967. Voting is now open.

MEMBER DOWDING: Hi, this is Dawn. Could you just say what the numbers represent again, please?

MS. QUINNONEZ: Absolutely. We're going to be voting for measure -- we're going to be voting. The criteria is 1 is high, 2 is moderate, 3 is low, and 4 is insufficient.

CO-CHAIR STILLE: Dawn, was that your question? Or was it the 100 point scale that the --

MEMBER DOWDING: No, that was my question. Thank you.

CO-CHAIR STILLE: Okay.

MS. QUINNONEZ: We're looking for one more vote.

And all votes are in. Voting is now closed. We have 6 percent voted for high, 11 percent for moderate, 72 percent for low, and 11 percent for insufficient.

CO-CHAIR STILLE: So this is on the scale measures. Okay. Do we continue to vote on the other sets then, right? Yes, we do. Okay. And the next set is global ratings.

MS. QUINNONEZ: Okay. Voting is now open for the importance to measure and report performance gaps for global ratings. Performance gap data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups or disparities in care. The criteria is 1 for high, 2 for moderate, 3 for low, and 4 for insufficient.

We're looking for two more votes.

And all votes are in. Voting is now closed. We have 0 percent for high, 56 percent moderate, 39 percent voted low, and 6 percent voted insufficient. And this is on global ratings.

CO-CHAIR STILLE: So that's consensus not reached, because it's 40 to 60 percent.

MS. QUINNONEZ: Would you like to continue?

CO-CHAIR STILLE: Okay. Yes. The next group is recommendations. Would you recommend X, Y, Z provider?

MS. QUINNONEZ: Okay. Importance to measure and report, performance gap for recommendations. The first criteria is high, second criteria is moderate, third criteria is low, and fourth is insufficient.

Okay. All votes are in. Voting is now closed. Zero percent for high, 67 percent moderate, 28 percent low, and 6 percent insufficient.

CO-CHAIR STILLE: Okay. So that group

passes. And the fourth group is unmet needs.

MS. QUINNONEZ: And the last group is unmet needs.

CO-CHAIR STILLE: That's the fourth of five. There's five groups and this is number four.

MS. QUINNONEZ: Got it. Unmet needs for the performance gap. The first option is high, second option is moderate, third option is low, and fourth option is insufficient.

Okay. All votes are in. Voting is now closed. We have, the results are 50 percent voted for high, 39 percent voted moderate, 11 percent voted low, and 0 percent voted insufficient.

CO-CHAIR STILLE: Okay. So that one passes. And then the last one is physical safety.

MS. QUINNONEZ: For our last category for safety for performance gap, our first option is to vote high, second option moderate, third option low, and fourth option insufficient.

All votes are in. Voting is now closed. We have 100 percent high.

CO-CHAIR STILLE: No, really? No.

MS. QUINNONEZ: Zero percent moderate obviously, 0 percent --

CO-CHAIR STILLE: No.

MS. QUINNONEZ: It's not right?

CO-CHAIR STILLE: It's not right. No.

MS. QUINNONEZ: What's not right? Let's try this again. One second. Okay. We're going to re-vote on our last category for safety. And it's going to be option number 1 high, option number 2 moderate, option number 3 low, and option 4 insufficient.

Voting is now closed. All votes are in. The results are 0 percent high, 22 percent moderate, 39 percent low, and 39 percent insufficient.

CO-CHAIR STILLE: All right, 39 and 39. Okay. So it does not pass. So we have some groups that passed. Sarah, have you been tallying those?

MS. SAMPSEL: I have. So we have the scale measures failed on evidence -- or on performance gap. The global measures, consensus not reached. The recommendation measures passed. The unmet needs measures passed. Physical safety measure did not pass.

CO-CHAIR STILLE: Okay.

MS. SAMPSEL: However, what we'll do though is, as we move forward, we'll still have the discussion about all of these for the benefit of the developers. But we'll only vote on those sets that either consensus was not reached, so the global measures and the recommendation and unmet needs measures.

CO-CHAIR STILLE: Okay.

MS. SAMPSEL: But we'll discuss all of them still.

CO-CHAIR STILLE: Yes, great. Okay. Good, thank you. Okay. Let's move on. Composite measure. What? No, never mind. Okay. All right, so let's move to discussion on reliability then. Do the primary discussants

want to talk about a couple things? Lee, Len, or Nicole, any thoughts about reliability adjusting?

There are two people with their cards up. Are you waiting? No? Katherine, are you waiting? You are waiting? Why don't you talk? That's fine.

MEMBER BEVANS: I think that, for me, the primary concern that I have about this measure relates to the specifications. Most notably the exclusion of people with cognitive limitations that, for the items that are being used, prohibit them or prevent them, perhaps, from providing reliable or valid scores.

And the reason, of course, I'm concerned about that is because, presumably, that makes up a large percentage of the population that you're targeting.

On the other hand, I might have missed it in the report. I wasn't sure if you had data on how many people were excluded because of those criteria, which I think will be critically important to know.

This, to me, also suggests a couple things. First, there's obviously been a lot of discussion about proxy reporters, which I think is important. But it's also been noted, and this relates to specifications as well, that there are typically quite a lot of disagreement, differences in opinions, between proxy reporters and people reporting on their own behalf.

So, getting at specifications, I think that moving forward -- I know that the data you presented were just for the self-report component of the measure. But moving forward, considering whether those, a proxy report measure and a self-report version of the measure, should even be considered comparable, or considered as two separate measures for consideration.

And the last thing I'll say about that, again, related also to reliability: did you do any cognitive testing with the items themselves to get a better understanding or, for those not excluded, how they understood the items, whether they were they meaningful, whether

they were well-understood?

DR. MALLERY: So, to answer the last question about cognitive testing, we did do two rounds of cognitive testing in both English and Spanish.

MEMBER MORT: Actually, we did three rounds.

DR. MALLERY: Three rounds. Sorry, Elizabeth.

MEMBER BEVANS: I'm bringing this up just to say that I think that -- and perhaps I missed it, there's a lot of material to look through. But, really, a clear description of that, that's so critically important for the target population -- and for everyone, really, but the target population in particular. Some more information about the results of those cognitive testing I think would go a long way to support the measure.

MEMBER MORT: Okay. Do you want written material or do you want a verbal description?

CO-CHAIR STILLE: We can follow up afterwards.

MEMBER MORT: Thanks.

CO-CHAIR STILLE: Okay. And let's see, Beth, did you have your --

MEMBER AVERBECK: One question on reliability. Were the survey methods tested compared to each other? Because you had both in person surveys and telephone surveys. And did one give you different results compared to the other? I know sometimes face-to-face compared to mail is a lot different, so I was just wondering about telephone compared to face-to-face. Thank you.

DR. MALLERY: We did compare, and the differences were, I think, significant on some measures but not on others. So it was not the same that you typically see in mail versus phone, perhaps because they were both in-person. And I can dig up those, you know, which measures were different.

But we do recommend that as a case mix

adjustor, the survey mode, to account for the differences.

CO-CHAIR STILLE: Lee and then Sherrie. And then I have a question.

CO-CHAIR PARTRIDGE: I had a quick question about the testing here. You said that you tested in 26 different programs and the total number of respondents was 2,300. That's a fairly small average number of respondents. How comfortable should we be -- or why should we be comfortable with numbers of that size?

I assume that's an average. I assume you had, in some programs, a much smaller number and a much larger number. But just to have a tiny bit of conversation, if we could.

DR. MALLERY: Sure. So I think there is a table that I can show, I believe in this package, that shows the numbers. But you're exactly right. There's a range, I think 150 is about the average. In some programs it was smaller. So some of the programs are smaller. So it was smaller.

And based off of this field test, we are recommending a larger sample size moving forward based on calculations for the -- in order to get a reliability of .7 for these measures moving forward, we're recommending an effective sample size of 400 for program moving forward. And we can talk about that a little bit more.

But I mean, that is what most of the programs are aiming for, for this next round. Maybe I'm over-speaking. But that's the guidance we've given for the next round. So on many of the programs, it is feasible.

CO-CHAIR PARTRIDGE: Yeah, I'm surprised. I actually think in some of these cases, you'd have a pretty small number of respondents. I mean, some of the incidence of this population is going to be quite small, right? For example -- but anyway. Thank you.

CO-CHAIR STILLE: And then Sherrie.

DR. JACKSON: I actually have statistics on that, if you'd like to hear them. This is for 2012, the data that's most recently

available in the HCBS 1915(c) waivers. Twenty-right percent have enrollees under 400; 30 percent between 400 and 3,000; and 41 percent between 3,000 and 50,000-plus.

So there's definitely a range. And the other thing to consider moving forward is, with the 2014 HCBS rule, CMS is allowing consolidation of waiver programs. So your size of your waiver programs are going to -- may grow over time.

Also with managed long-term services and supports, that, essentially, for many programs, is consolidation. So the size of your programs over time is likely to grow.

CO-CHAIR STILLE: Sherrie, you had a point?

MEMBER KAPLAN: Let me follow up because I'm old and I forget things like crazy. So let me follow up on that point first. The issue of, you know, if you're sampling, if you fix the sample size, you're going to get a very much different, and probably more stable,

estimate of the smaller programs than you are of the larger programs, depending on how you do that sampling.

Because you're going to have to be careful about who you're representing in that. You know, it's really going to have a very strong randomization and sort of random sampling issue to get a stable estimate.

So if you fix the sample size, it raises more of a validity than a reliability question. But you're going to estimate some of these programs with half of their population and others with a very small portion.

So the other issue is that some of your scales have fairly good precision, in terms of at the patient level anyway. And I'm looking at Exhibit 2. I forget what page it's on. But the three scales that are really fairly strong are staff reliability and helpfulness, staff listening and communication, and case manager helpfulness.

And then it drops to choosing the

services that matter to you. And the reason I'm going into this I'll come back to in a minute. And that actually drops pretty low. On the other hand, that only has two measures in it, two items in that measure. And that's exactly what you would expect given the way this formula is constructed.

And then down to, you know, sort of the personal safety and respect, that's got three items in it. And again, you would get -- you would expect that because it has fewer items in it.

So my first question is, have you a done a look, or could we ask you to do a look, at a Spearman-Brown prophecy formula, to say this is an important construct to us but I don't really have enough items in that construct and it's not robust enough.

It may actually be really good in discriminating facilities one from another, but because I don't even have very good reliability, I can't tell that. So do you need more sample

size? Or do you need more items? And the Spearman-Brown prophecy formula data would give you that. So that was question number one.

And question number two is, the inter-unit reliability issues are problematic for some of these unmet needs measures. And specifically the unmet need in toileting. It doesn't have very much variability at all. But it's got crummy IUR, it has terrible IURs. But, you know, some of them, unmet need for medication, have pretty good and pretty robust IURs.

So, you know, is there a possibility going forward that you can look at maybe the intraclass correlation coefficients at some of these units being compared and look at within, versus, between unit comparisons and see if there's some gain that you could, possibly, with an increase in precision of these measures, that you could actually look at.

Or -- and this is horrible for NQF and I apologize, Sarah -- the possibility of creating a composite out of this where you're creating a

composite out of unmet needs. It doesn't prevent you from drilling down to individual item level but it could help you in the precision.

CO-CHAIR STILLE: Great, thanks. Sam?

MEMBER BIERNER: I want to comment specifically on the table that's just above 1.6 where you list the number of returned surveys. There were only 92 brain-injured patients surveyed that returned surveys. I don't know how many you actually surveyed, but only 92 counted in this sampling.

So it seems like you've under-represented TBI for purposes for drawing any conclusions about them. And I wanted to see what your comments were about that and how you would correct it.

There's a table that shows number of total returned surveys. From Maryland, there was zero that was thrown out. And then 72 from Minnesota and 20 from New Hampshire in that table.

DR. MALLERY: So in that table it

might be a little bit misleading. Because I think that's just within -- or maybe I'm -- in Exhibit 1 that goes through the states?

MEMBER BIERNER: Yes.

DR. MALLERY: Okay. So there were actually, overall, 262 beneficiaries with traumatic brain injury in our sample. So I'm thinking that maybe that that is one TBI program within one state that you might be looking at.

MEMBER BIERNER: Right. Yes, I see that in the top part of the table. But there's still a very small sampling. You said you'd need a minimum of 400. So you less than 250.

DR. MALLERY: That's correct.

CO-CHAIR STILLE: Okay. Good. Great discussion. We are now going to vote on reliability for the three remaining sets, since there seems to be some differences in numbers between the sets, with global rating, recommend, and unmet need.

MS. QUINNONEZ: Okay. Give me one second.

(Pause.)

CO-CHAIR STILLE: Desi, let us know when you're ready.

Do a hand vote? Okay. We can do a hand vote if that's easier.

MS. QUINNONEZ: Okay. Voting is now open for scientific acceptability and reliability, including precise specifications and testing, for the first group, which is global ratings. And the options are number 1 high, number 2 moderate, number 3 low, and 4 insufficient. Voting is now open.

Okay. We're looking for four more votes. Two more. Okay. All votes are in and voting is now closed. The results are zero percent high, 39 percent moderate, 44 percent low, and 18 percent insufficient -- 17 percent insufficient.

CO-CHAIR STILLE: Okay. So that's for the global rating group, right? Okay. And that fails. Okay.

MS. QUINNONEZ: Okay. Voting is now

open for reliability, including precise specifications and testing, for the recommendations group. And the options are number 1 high, number 2 moderate, number 3 low, and number 4 insufficient.

All votes are in and voting is now closed. The results read 0 percent for high, 22 percent moderate, 67 percent low, and 11 percent insufficient.

CO-CHAIR STILLE: Okay. So that fails on reliability. The last one is unmet needs.

MS. QUINNONEZ: Voting is now open for scientific acceptability and reliability, including precise specifications and testing, of unmet needs. The options are number 1 high, number 2 moderate, number 3 low, and number 4 insufficient.

All votes are in and voting is now closed. The results are -- no. We'll do this one over. Give me one second.

MS. THEBERGE: Dawn, I don't have your vote yet.

MS. QUINNONEZ: We will re-vote again for reliability, including precise specifications and testing, for unmet needs. Number 1 high, option number 2 moderate, option number 3 low, and option number 4 insufficient.

All votes are in and voting is now closed. The results read 6 percent high, 11 percent moderate, 67 percent low, and 17 percent insufficient.

CO-CHAIR STILLE: Okay. Absolutely right. So we're done with voting but we'd like to continue the discussion to be as helpful as possible to the developers as we can give them some recommendations for where to go next.

I'm just trying to recap what I remember so far. One of the points was basically more numbers for more populations. One of the points was really looking hard at proxy reports. And both doing more of them and also differences with self-report. There was one other big point and I can't remember. Anyone?

MEMBER KAPLAN: Intraclass correlation

coefficients, between unit differences.

MEMBER BEVANS: Maybe showing the evidence from cognitive testing for understandability of items.

CO-CHAIR STILLE: The other thing that just came to mind is, especially with some of the ones that seem to have the ceiling effect, some quality improvement folks do what's called a top-box reporting system. And, you know, the validity of that, it kind of depends on what you're trying to say. But that might be useful to look at, too. Lisa?

MEMBER SUTER: First of all, I'd like to echo Sherrie's recommendation for looking at the Spearman-Brown prophecy formula. I thought that was a great suggestion. And along those lines, I think understanding the feasibility of getting to your optimal sample size of 400 would be really helpful, since your sampling is much below that level. So, understanding what the burden on providers is to get to that level.

I'd also love to see information about

how long it takes patients or family members to fill this out. So, the burden on the individual patients or caregivers.

And the final thing is, can you describe a little bit more -- maybe not now, but in the materials -- what the actual patient input was to development of the questions. There's a notation, and I may have missed the more detailed information, that this population demonstrated that these were important. But I wasn't sure whether they were actually involved in the development.

CO-CHAIR STILLE: Okay. Peter, and then Lee.

MEMBER THOMAS: So we had a motion on the table at the beginning of this discussion that we never acted on. And this may be inappropriate; let me just put it on the table. You've heard this whole discussion and you've gotten the comments. Is this something that you can "cure" in two months? Or would you rather have a deferral of this whole discussion for some

longer period of time and come back to the table?

DR. MALLERY: Many of the things you've asked for we could pull together in a few days -- I mean, it's stuff we've already done. I mean the cognitive testing, how patients were involved -- they certainly were. And we could have a report on that, we could pull that together very quickly.

More data is a different story. So I think that's what we've have to talk about. Because there is another round -- but just more data is a whole other story.

MS. SAMPSEL: Chris?

MEMBER KAPLAN: Sorry, one more thing. If you're going to do the exercise of going back through the data, you know, factor analyzing, you've got enough subjects to do a factor analysis, would help how multidimensional this construct is. Because you're losing precision if you're doing these fine grain things.

It actually may help you more if you're doing a bigger composite. So, you know,

you're kind of crunching things into -- and that's why the Spearman-Brown prophecy formula early on would help you understand, am I needing more measures or more subjects per measure?

And, you know, doing that early and then giving us a factor analysis -- or doing a factor analysis and saying, geez, I've cut this too finely, the factor analysis isn't supporting it.

DR. MALLERY: I don't know if I should respond. But, I mean, all of these measures are based on factor analysis. So we did do that. So there was confirmatory and exploratory. And I think that that is probably written up a little bit here. But happy to give more detail about that.

MEMBER KAPLAN: Yeah, that data would have helped. But it depends on how you do those rotations, because these things are notoriously -- so an oblique versus a varimax rotation, for example, would help you with that issue, on how multidimensional it is.

MS. SAMPSEL: I guess my comment is just going to be to go back to -- I don't think we need to make a decision on voting on a deferment today. I think we need to go back and give you guys some time and be clear on what we want to do and we can have another discussion.

And it might make sense actually for us to mimic some of your forms so that we have five submissions, and so that we can deal with them separately and make it a little bit of a cleaner process as well. Because I think some of these are being held hostage, perhaps. Although, you know, they all failed when we got down to it.

So let's see what kind of additional data and we can bring it back for the post comment.

CO-CHAIR STILLE: And then Linda.

MEMBER MELILLO: I would just say, if you're going to include employment questions, to bring them, as well, if you're thinking of adding them to this grouping.

DR. MALLERY: I don't think we were

considering putting them up for endorsement. But they are available on the survey. So I think that's important to note, that for use they'll be available.

MEMBER MELILLO: Okay.

CO-CHAIR STILLE: Okay. Others? Okay. Well, thank you very much. I'm sorry this was really hard, but it was a good process. And I'm sure it will win eventually. Okay.

(Pause.)

MS. SAMPSEL: Operator, is Nick Castle on the phone?

OPERATOR: Yes, he is.

MS. SAMPSEL: Okay. Can you just make sure he has an open line?

OPERATOR: His line is open.

CO-CHAIR PARTRIDGE: All right, we're going to switch over to -- and I think we'll probably pretty easily be able to talk about them as a group. Three measures, all related to discharge -- sorry. I think I'm still tongue tied from the last one.

First, 2614 is a measure of taking at discharge from somebody leaving a short stay. 2615 is long stay resident who is presumably not leaving. And 2616, long stay, but this time the family is the respondent and not the patient. So we'll turn it over to our friends from AHCA.

MR. MULLER: Very good. So I will keep this brief. So the three CoreQ measures, we really wanted a short set of parsimonious and effectiveness, experience of care measures. And brought Nick Castle, who's on the line, from the University of Pittsburgh in to develop them.

For those who don't know, Nick was probably the core mind in developing the nursing home CAHPS measures. And essentially, this is a short form spiritual descendent of nursing home CAHPS.

To build it, we started with a large set of items and did rounds and rounds of exploratory factor analysis to distill down to a minimum set of measures that captured the bulk of the signal picked up in the larger measure, large

instrument and larger measure.

And then we intensively tested the kind of agreement with patients agreeing that they are important measures and clear measures and all of that. And it was backed up very nicely.

And so, the three measures are short stay discharge, long stay resident, and long stay family. And they're all based on a one to five, poor, average, good, very good, and excellent scale.

The questions, to give you a flavor of the definition of the measure are one, in recommending this facility to your friends and family, how would you rate it overall? Two, overall how would you rate the staff? Three, how would you rate the care you receive? And then just for the short stay discharge measure, how would you rate how well your discharge needs were met? So very much kind of an experience of care measure that parallels a little bit the five star branches as well. The measures then, the percent

of the responses who averaged, whose answers averaged good or better.

The last thing I'll say -- I'll not get into any of the other technical details. But we have ten customer satisfaction vendors who have all invested to incorporate the four questions into their instruments. And this is, I don't know what portion of the market for the skilled nursing market for customer satisfaction surveys is not captured by these guys. But I'm guessing that it's very small.

And so, what this represents is ten very serious professional survey firms effectively, you know, a free market sense endorsing the measure for it's sort of appropriateness and all of that as, eventually, kind of a uniform national standard for customer satisfaction measurement in SNFs.

So it's an important thing there. And it's currently being administered and it's currently being collected. And in our AHCA quality initiative, in some of our tools we're

beginning to collect the measurement themselves. And this will just sort of grow over time. And with that, the discussion.

CO-CHAIR PARTRIDGE: Our lead discussant group has gotten a little smaller. We've lost David. Now I've lost my cheat sheet. Sherrie, I think you're one on these. Am I right? And Peter is the other. Okay. Either one of you want to go first? Peter, do you want to go first? Good.

MEMBER THOMAS: Yes. I'll just say a couple of things. And again, you don't really want to go into evidence or anything. You just want some overarching comments? So this is a patient reported outcome measure and it's obviously a major area.

Skilled nursing care, skilled nursing facilities provide care to millions of Americans each year. I agreed with the -- well I was already into evidence here. Hold on. I had some real questions about the exclusions because I felt that some of them might impact the overall

satisfaction scores in a significant way. But I can go into that in more depth when it's time to do that, during the evidence discussion.

But it's obviously a very important measure for those people who go to a nursing home or a skilled nursing facility and don't stay indefinitely, don't stay a long term period of time. And that's apparently more and more folks, the more rehabilitation is being provided in those settings. And that's obviously a very important thing to measure, is them being satisfied and being ultimately discharged back into the community.

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: I had some -- first of all, I support all of Peter's comments. And then I had some questions, more questions than anything else to the developers. I agreed that the exclusions actually may limit the generalizability to a small proportion of facility nursing home patients.

And then the consistency of the

implementation across facilities actually could be compromised by the low response rate. So I was a little bit concerned about the response rate.

Second thing I was concerned about was there was test retest reliability but they give one month as the testing interval. And they give percent agreement but not intra-rater reliability at the patient level. And they didn't give ICCs at the facility level, at least I couldn't find it.

The other thing I was, a few other things. The pre-specify, when you do boot strapping, you can't really interpret it without pre-specifying a minimally important difference. So I would like to hear something about what you considered a minimally important difference.

And the sample size varied from 20 to 196 per facility. So the stability of estimates across facilities is going to, that's going to affect the stability of those estimates.

And then I didn't get the item to

total correlation coefficients for the 22 items to support the item choice thing and your elimination of the items.

And in the validity testing, although your factor analysis looks like it supports a single dimension, you don't say if you did any varimax or oblique rotations to kick the tires on the multi-dimensionality. And so I was interested in hearing your responses on that.

And finally, the validity concerns I had were like because, although you note that there are these correlations, some of them are significant with the validity variables you chose but a number of them are very low. And many of them are not significant. And so, then I was wondering what we were going to conclude about the, your take of what that means.

CO-CHAIR PARTRIDGE: That's kind of a laundry list. So I'm not sure that we can ask you to respond to all of them right this second. But if you want to tackle, you know, a couple that seem -- either we can knock it off fast or

it's --

DR. GIFFORD: I assume with the exclusions people are concerned with the cognitive impairment piece for it. Because the rest are really minor. And I would think clinically, seeing that makes sense. People die, can't really survey them. And we're really split out making sure it was the voice of the patient or the voice of the family. We would not combine them. So we exclude them that way.

The cognitive impairment one has, you know, been really a difficult one in satisfaction in general in this setting. And I think we ended up setting on, using the BIMS score that Salida developed for MDS. So it was uniform to every nursing home resident out there. So we had a standardized way of assessing that. It wouldn't allow people to game the system.

It's currently, something like that is used currently pretty much across the board in most nursing home satisfactions that are out there thought not everyone can quite do it. We

set it at the level of where we did because it's pretty moderate to severe. It's unclear anyone would have a short term memory to be able to recall anything overall.

It mainly was -- frankly we don't care. You can send it to them. But the chance that they'll fill it out or not have someone else assist them in filling it out, then their results would be excluded anyway because we don't allow proxy individuals. Then you just, you dry the sample, your response rates become a problem. And so, it was also just to help with the response rates frankly.

MEMBER KAPLAN: May I ask a follow up question on that? Because the response rates don't look great to begin with. And so, I know this is a vulnerable population and it's difficult to reach and all that stuff.

On the other hand, when you've got that limited amount of response rate, can you give us a sense of whether those, the patient characteristics match the profiles of the

facilities? Or anything to help us see if that's a generalizable sample you got?

DR. GIFFORD: I was a part of that and I don't know if we did the latter part or not. But I know that one of the things we did is we talked, we talked a lot with a lot of the, as James said, a lot of the survey vendors out there. And pretty much across the board, most everyone right now is accepting somewhere around 15 to 20 percent which we thought -- yes, that was also my response too.

And so we boosted it up. And then we really pushed it. We're allowing right now, in sort of initiative, anyone to have any response rate. But when we calculate it or do any reporting or do any reporting, we want a minimum response rate.

I mean, if you look at even the states that are publicly reporting satisfaction out there and there's four or five of them, we talked to those vendors in those states and they're all accepting really low response rates. So we ended

up deciding to go with that from a pure usability standpoint.

Because while I completely agree with you, I'd love to see a higher response rate, even in the states where they're doing public reporting, you would essentially probably go down to about 10 to 15 percent of SNFs in the country having an adequate response rate. Then I'll turn it over to James for the other analysis.

MR. MULLER: So for the technical questions, do you want to step through them one by one?

MS. SAMPSEL: Actually, can we hold those until we get to scientific acceptability please?

MR. MULLER: All right.

MEMBER THOMAS: First do the exclusions for a second.

MR. MULLER: Sure.

MEMBER THOMAS: Because my concerns were a little bit broader than just the cognitive --

MR. MULLER: Okay.

MEMBER THOMAS: -- exclusion. And this is going to sound maybe a little tongue in cheek. But some would argue that the person that dies is the least satisfied with their care.

MR. MULLER: Yes.

MEMBER THOMAS: And I'm concerned about also the readmissions issue, about going to another facility. It might have nothing to do with the care being provided in the SNF. But it might have something to do with the care being provided in the SNF. And if it is because of some of the things that are being measured here, competency of the staff, you know, responsiveness of the management, if those are in fact contributing factors to why that person has to go to another facility whether they're being readmitted to the acute care hospital or to a higher intensity level of care like an IRF or if they die.

Those are perhaps, you might argue, the most egregious cases and they're being pulled

out of the sample. So can you just address that?

DR. GIFFORD: We talked a lot about it. We actually would prefer to see a hospice level or end of life, you know, some of the survey stuff that Joan Teno's down at Brown. Because I think she's also shown that this group is the experience and the questions -- it's not even clear in the focus groups in cognitive testing whether these are the appropriate questions to ask for people who are dying.

The experience of the provider community has been sending any sort of survey questionnaire to family members of people who have died has often been very offensive. And so, it was felt to be out of respect of them to collect the data a different way. And they are excluded for that reason as well.

MEMBER THOMAS: That's fair. I guess where I'm going is maybe something that's in politic to raise, I'm just trying to understand. If you've got a SNF that's not up to par on quality and they're actually contributing to the,

some of the things that are happening to these people who are then being excluded from a quality measure, isn't that biasing the data in a favorable way?

MR. MULLER: I think if the SNF is systematically poor, it'll be picked up as well by the patients who aren't being excluded. So that will be driven down by that signal as well.

DR. GIFFORD: We also have a panel of other measures that we're held accountable to that are using -- so we have discharge to community. We have rehospitalization rates. We have more functional improvements to come. So we have a whole set of other metrics that are being held accountable to that provide an opportunity to sort of balance, I think, if we're discharging inappropriately or having other problems that are out there.

CO-CHAIR PARTRIDGE: Before we move on, Deb you are back on the line with us. Is that right?

MEMBER SALIBA: Yes I'm back.

CO-CHAIR PARTRIDGE: And you're one of the discussants here. So maybe we should give you an opportunity.

MEMBER SALIBA: Yes. So I certainly agree with the premise that obtaining resident satisfaction measures is really important. And that this is a gap in the current measurement approach. So I think it's great that this is being done.

I was a little bit unclear if the CoreQ analysis that we're seeing is having received a full survey and then answered these summary items. Or if it's just answering these summary items. It seems like in a lot of satisfaction surveys, you tend to have the more granular items leading up to then asking these more summative kinds of scores.

And so, it was a little unclear to me if the testing results that you're showing us are just, you went in and just asked these three very summary questions which seem to be a little bit challenging in terms of providing an actionable

road map. Or if you're going in and doing the longer survey and then only scoring the last, the summative three questions.

The rest of my questions were about the more, along the lines of the evidence, the correlations and things like that. But I did want to ask about that question.

MR. MULLER: I mean -- and Nick can speak to some of this as well. There were different rounds of development. And some of the earlier rounds of development needed the full set of questions so that we could whittle things down. And I think the current form that he's collecting out there is distilled down to the minimal set.

Some of the vendors that are doing it have integrated into their own instruments with other questions in there and that sort of thing. So it's really a mixture of different things.

MEMBER SALIBA: Another thing was the choice of the response scale, poor, average, good, very good, and excellent. So it seems

pretty heavily weighted towards positive responses. And lumping the good, very good, and excellent together would seem to weight more in the direction of positive response.

And I was interested in how you came at that particular response scale.

DR. GIFFORD: Well we actually, and Nick had commented, we did focus groups and cognitive testing of different response scales from ten point down to four point Likert scales. And this actually is -- I mean they all worked reasonably well.

And Nick can talk about some of the work he's done with some of the CAHPS. But he found that most of the CAHPS surveys, no one answered below five or six. And so, that's one reason we collapsed, in general, in satisfaction, when you're asking poor or fair, it's bad.

To your point, we actually tested five or six different ways to calculate the measure. And I was thinking that we could find a better way of capturing it. And it turned out it did

not matter how you collapsed this together. You just ended up with different satisfaction scores but relative ranking was no different. The only thing that really different was if you just took the top rating of whatever scale you used, that was a fundamentally different measure.

But five other different ways of averaging them together, you just ended up with a different national average or group average. But it didn't vary at all.

And then I think, back to your earlier point, the early round testing was all the questions. And Sherrie, this may go to your points too. Our goal was to get a parsimonious measure for accountability purposes. This was not for quality improvement purposes. Let me make that really clear.

I mean, if you were doing quality improvement purposes, you put food in this question. It doesn't matter whether you ask food. Once you ask these three questions and add them together, your scores don't change. Your

quality improvement and input is tremendous.

And so, this was one reason we also recommended that the current vendors -- and why they all support it. It gets added as a block set of questions by themselves with the standard -- because they all used different Likert scales and do everything differently. And so it's really been important.

The other is -- Nick do you want to talk about? We asked, would you recommend to a friend? And how satisfied were you overall? And we put it at the beginning. And we put one at the beginning and one at the end and we flipped it beginning to the end. And it is absolutely no different.

DR. CASTLE: Yes, I was about to try and comment on that. And I agree with Deb's point. And we did a decent amount of testing on this because we had the opinion that if you have these aggregate items first, you do think aggregate. And then you answer the more granular questions -- is that going to have an effect? As

opposed to answering the very granular questions that may then impact your overall opinion of the facility at the end.

So in the initial testing, we had the questions at the front and at the back. And in several thousand surveys, we found really very little difference. But again, I agree with Deb. We expected to find something, we didn't.

But what difference we did find is if you have the block first of aggregate items, we get all of the items answered and we get a very, very good, complete response rate which of course we need for calculating the score. So it was advantageous to have the items blocked at the beginning.

DR. GIFFORD: So we originally calculated from -- I forgot. Nick, how many was it? Like 20 or some odd questions. But then the --

DR. CASTLE: 24.

DR. GIFFORD: 24 but then the subsequent reliability and validity of the large

sample size was done just on these three questions or four questions for the discharge. And they were done with, either block by vendor that was out there added within the survey.

CO-CHAIR PARTRIDGE: Are we ready to move on to evidence? Sorry Becky?

MEMBER BRADLEY: Just as a point of clarification, I guess your previous answer was that this is not in addition to the CAHPS surveys that are being done. This would be instead of for the patients who have a short stay of 100 days or less in a skilled nursing facility. So it would replace a CAHPS survey for this particular post-acute setting?

I guess I'm -- we spent a lot of time at our last meeting talking about how CMS is trying to standardized instruments across post-acute. So I guess I'm trying to figure out where this survey would play out in the post-acute setting. And how it would be compared to other post-acute settings going forward. Or is it just unique and by itself?

MR. MULLER: A little bit of both. I mean, it's intended, it wouldn't make sense to do both the nursing home CAHPS survey and this survey because they will correlate very highly. Because the additional items, as Giff sort of pointed out, don't really -- it's picking up the signal. And so there's a redundancy between the two.

This is a shorter, sort of more parsimonious survey that will get a higher response rate because of that. So I guess one could do both. But I don't know why one would.

DR. GIFFORD: And nursing home CAHPS is not used right now. No one's using it. There's a handful using it for discharge. But other than that, no one is using it for long stay or anything else. And then CMS just went through the endorsement application. So it's not even an endorsed measure anymore.

So right now there is no standard metric out there. And this started because we made imprudent satisfaction as a primary focus of

our national quality initiative. And when we went out there and looked at everything being done, it's all over the map.

And so, the feeling was we needed a standardized way to do this. And that's why we worked with Nick and all the vendors. And everyone's been really excited because it's the first time there's a standardized metric that you can compare across everyone.

MEMBER BRADLEY: Across --

DR. GIFFORD: Across SNFs, yes.

MEMBER BRADLEY: But not across the post-acute?

DR. GIFFORD: No. We developed it for -- we don't have it to bring forth. We developed it for assisted living too. And we've tested it and we've done a lot of collecting. The problem with it, we don't know how to test the validity of it because there aren't good measures of other external measures for validity out there.

MEMBER BRADLEY: Okay. Thank you.

CO-CHAIR PARTRIDGE: Can we move on to

evidence? Any comments or questions from either our discussants or the rest of our membership on evidence starting with 2614 which is the CoreQ for short stay discharge? Sherrie?

MEMBER KAPLAN: Can I just say something positive?

CO-CHAIR PARTRIDGE: Yes.

MEMBER KAPLAN: So yes, one, I really liked there was a conceptual framework that was provided at the beginning that linked the measure with other things that were going on. Second thing I liked, they really did cite our literature review kind of as a really, a more comprehensive literature review that documented the evidence for a lot of variation.

And then they showed that there was a considerable -- although it's still favorably skewed, there was a gap, there was a demonstrated gap. And the interquartile range was like 13 points which is unusual in these satisfaction measures because you get them always bumped up to the top. So I did think there was strong

evidence provided both for the gap and for disparities by age and gender.

CO-CHAIR PARTRIDGE: Thank you. Linda?

MEMBER MELILLO: Yes. I would just like to say there is no required reporting or measurement in the SNF population currently. And I appreciate it's short. Because especially in that population, which tends to be very sparse in terms of staffing, that would be much more doable than one of the lengthy surveys.

MR. MULLER: Yes. We all hate paperwork.

CO-CHAIR PARTRIDGE: So are we ready to --

DR. GIFFORD: There's like five states that do public reporting in the nursing home but none national. And three other states have incorporated this into their questionnaire.

MEMBER MELILLO: Massachusetts has a state survey that they do annually. But yes, but again, it's very length. So this would be

preferable by far.

CO-CHAIR PARTRIDGE: Desi?

MS. QUINNONEZ: Okay. We are now voting on Measure 2614 CoreQ short stay discharge measure on evidence, the rational supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service. The voting options are number one for yes and number two for no. Voting is now open.

Looking for two more votes. Voting, all votes are in and voting is now closed. The results read 94 percent voted yes and 6 percent voted no.

CO-CHAIR PARTRIDGE: Okay. Moving on. I think Sherrie has already touched briefly on gap. Any comments specifically on this particular aspect? Deb, chime in here where you wish. And Lisa, you were, I can see your card.

MEMBER SUTER: So yes, could I just clarify, are we only voting on one measure through all criteria?

CO-CHAIR PARTRIDGE: Yes. I'm sorry. We are going to go through -- because they are slightly different, we are going to go through. And we also happen to have a discussion of the other two. So I'm sorry, back to voting on -- Desi.

MS. QUINNONEZ: Are we ready for gaps? Okay. Voting is now open for Measure 2614 for gaps, performance gap data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups. Option number one is high, option number two is moderate, option number three is low, and option number four is insufficient.

Waiting for one more vote. Okay. All votes are in and voting is now closed. The results read 39 percent voted high, 56 percent voted moderate, 6 percent voted low, zero percent voted insufficient.

CO-CHAIR PARTRIDGE: Okay. Moving on to scientific acceptability. First one, our

friend reliability. Floor is open for comments, concerns. Sherrie?

MEMBER KAPLAN: This is for the developers, do you have a sense of, do you have the intraunit reliability, the intraclass correlation coefficients at the facility level for these, for this measure?

MR. MULLER: Nick?

DR. CASTLE: Actually I do not. But when you say interunit, do you mean nursing home unit or --

MEMBER KAPLAN: Yes, facility level. Whatever unit is being compared. If it's the facility that's being compared, then evidence that it's --

DR. CASTLE: Oh so this was at just the facility level. I would love to do it at the unit level. But we had names and addresses of residents, clearly. But we don't know which units they were on.

MEMBER KAPLAN: I guess I'm asking the facility level with intraclass correlation

coefficients showing that it was more between than within unit facility variability.

DR. CASTLE: No I do not. I was actually trying to look that up right now when you mentioned it earlier. I haven't found that yet, no.

MEMBER KAPLAN: And then one quick question on battery reduction. So in this battery reduction exercises -- because these things are going to be really highly correlated with each other, they tend to be. To get from 22 items to 4 items, this is back to the Spearman-Brown prophecy formulation, you're going to lose some precision when you shrink the number of items. But you know, if you're losing variation -- I didn't get the item to total correlation coefficients for the 22 items.

So if you went through that data, as you probably already did, how did you choose the top performers there? Because if these all have kind of a fairly strong relationship to the total body of items included, how did you pick which

ones to --

MR. MULLER: Again, a question for Nick.

DR. CASTLE: So some of this, as you might have guessed by the application, that this was extremely iterative. We did this many hundreds of times. But some of it was when we dropped our doors with the highest correlation. Then we continued to drop out that way. We would go partially down and then add some back in to see, as you're saying, that we hadn't actually mis-dropped something.

So it was almost a reverse process from what folks often do. We were trying to get items that almost were capturing the most satisfaction information that wasn't within another item. So if the items really correlated very, very, very highly, it made sense to drop one of those items. But once we had, later on we would put it back in until we were comfortable with dropping the individual items.

Now the other thing that was going on

with this is with the three instruments, we were also trying to do, to get items that were similar in the three instruments also. So it was almost a family of instruments.

DR. GIFFORD: The other thing that was going on at the same time is we were trying to look for a personalized number for a measure. So it wasn't that we were just trying to get a scale. Because they perform well as a full scale. And you can pick, sort of different numbers. But it was, what was the minimum number to get an aggregate measure to hold to assess the facilities overall satisfaction?

DR. CASTLE: Actually I think Giff just answered that way better than I did. If I answered a different way, I'd say, so if we had three or four items in there. Let's say we have four and we would calculate the score with those four items. We got somewhere, depending on which instrument we're talking about, 84 percent to 89 percent or in some cases 91 percent of the score we would get with the full instrument, or with a

score using the CAHPS instrument which seems to indicate that we're capturing overall satisfaction. And if we added another item in, we might go up to 86 percent rather than 84 percent.

So the addition of the item wasn't getting, wasn't capturing much for us.

MEMBER KAPLAN: I get it at the patient level. What I was looking for was at the between facility variance versus the within facility variance. And when you've got only 20 subjects for some of these facilities, what you've got to go back, what it would be really good to go back and do is look at the Spearman-Brown prophecy formula to see how close to the bone are you getting with four items. Especially as you get sample sizes that shrink that far.

So you know, if you're getting too much noise, then you're going to get a lot more within facility variability that's going to mess up your ability to discriminate between facilities. So that's, it's related to that

question.

It doesn't necessarily -- I don't Sarah. It doesn't make the bar right now of the things that NQF is going to ask for. But it may make the next round. Is that right? Or is this a legitimate question for this round?

MEMBER SAMPSEL: I don't think so. I mean, because basically what we would like you to get out is have they provided the level that we have, you know, kind of -- go back to the algorithm, Sherrie, that you questioned earlier. And that, you know, what we've already communicated to the developers is that we want facility level testing in order to be able to interpret the measurement score.

So anything beyond that is something we want to see in the future. But we haven't stated it.

MEMBER KAPLAN: This would be within then, what you just said. It would be asking for between facility, evidence that there is between facility variation that's not a function of

within facility variation. And my only concern is, when you're getting that small number of items, that you could noise up your variable and start causing yourself trouble.

CO-CHAIR PARTRIDGE: I see Peter's hand up.

MEMBER THOMAS: Real quick, there's no risk adjustment here, right?

MR. MULLER: Correct.

MEMBER THOMAS: And is that because the four questions were, you didn't feel that you needed to do risk adjustment when you're asking those four patient reported outcome measures?

MR. MULLER: It's really the concept. So we thought about is there like a clinical driver that would make one person more satisfied than other? If it's appropriate to control that. And we couldn't think of anything.

Sociodemographic characteristics, so race or something like that there's, I think there is a small correlation with race. But again, it's like, it feels very inappropriate to,

if one race is systematically more satisfied than another, to control that difference.

So we sort of could not find a reasonable set of factors or any factor to adjust for.

MEMBER THOMAS: You've also got --

DR. CASTLE: I would answer that we searched, as James said. We did look at SES. And whatever factors we had available, we searched the scores to see if any kind of risk adjustment was appropriate. And it's actually also not standard to risk adjust for most satisfaction surveys in long term care. But we did look into it and it didn't seem to make sense.

MEMBER KAPLAN: Following up on that, in 1B you do note age and gender differences. Is that what you're excluding for risk adjustment?

DR. GIFFORD: We don't do any risk adjustment. I mean there are age and gender differences. Part of it also is the general membership of, felt that they want to be held

accountable for who they take care of and not risk adjust away differences.

MEMBER KAPLAN: But no, older people are more satisfied and women are less satisfied?

DR. GIFFORD: Yes.

MEMBER KAPLAN: Independent of the care they get?

DR. GIFFORD: Yes.

MEMBER KAPLAN: Okay.

CO-CHAIR PARTRIDGE: Are we ready to move on to voting? Sarah's nodding. Everybody else? Are we all awake? Okay. Desi?

MS. QUINNONEZ: Yes. Voting is now open for Measure 2614 on reliability including precise specifications and testing. Option number one is high, option number two is moderate, option number three is low, and option number four is insufficient.

We're looking for one more vote. All votes are in. Voting is now closed. The results are 33 percent voted high, 44 percent voted moderate, 22 percent voted low, and 0 percent for

insufficient.

CO-CHAIR PARTRIDGE: Variability, again, floor is open for comments. Sherrie is that card up or down?

MEMBER KAPLAN: It's up. And I only have one question. My question is on Table 2B2.3.J on Page 61, you got a variable going in the wrong direction with respect to readmissions. And because most of your variable evidence, the validity evidence is modest at very best and in many cases nonsignificant, what do you make of the one that's going in the wrong direction?

MR. MULLER: Which page was that?

MEMBER KAPLAN: Page 61.

DR. GIFFORD: I remember this.

MR. MULLER: Yes.

MEMBER KAPLAN: It's a strong positive. It's one of your strongest correlations and it said it in the wrong direction.

DR. GIFFORD: I remember that. And we went back and reexamined that to make sure we

didn't enter it in wrong. We thought it was a typo too. We elected not to make up an excuse because we just couldn't figure how to do it.

But you know, with any of these correlations, we could always change our hypothesis after the fact to make them fit. And we elected not to. We felt the preponderance was in the way we wanted. And I have no idea Sherrie. I really do not know.

MEMBER KAPLAN: If they had a done a Fisher's exact test and counted up the number signs that were in the correct direction, this wouldn't have screwed you up, wouldn't have messed you up. But on the other hand --

DR. GIFFORD: That's kind of what we did.

MEMBER KAPLAN: I'm just very curious. It's the highest correlation.

DR. GIFFORD: I know.

MEMBER KAPLAN: And it's going in the opposite direction.

DR. GIFFORD: You should have seen the

reaction. Actually we didn't even pick it up on the first draft of this. We talked about how this was the strongest correlation and everything. And then someone on the review said, but it's the wrong direction. And we had to go back and rewrite. We just elected to -- I don't know. I don't know how to explain it. It's just one of those things.

MEMBER MELILLO: I may have an answer for you. Pressure ulcers, is that what we're talking about?

DR. GIFFORD: No, rehospitalization.

MEMBER MELILLO: Oh, nevermind then. Oh okay.

MEMBER SALIBA: So this is Deb. My only comment was just that the Fleishman that you point to as indicating understanding is more just a reading level test than understanding. You may have also done some cognitive testing with the items to see if people are answering the questions we think they're answering. But the test that you used is probably just a first step

and not really a test of comprehension.

DR. CASTLE: I can answer that one Deb. Yes, we did cognitive testing with family members, with residents, and with short stay residents. We went to facilities in Pittsburgh. And with the cognitive testing, I think we had more than 100 in each group by the time we were finished.

And we did it the classic way of reading questions and having them respond back on what they thought we were asking and if we could ask it differently and what they understood with words.

The Fleishman was put in there because it also passed that test. But the cognitive testing was exactly as you describe, the classic way of doing it.

MEMBER SALIBA: Great. I figured you probably had. Thank you.

DR. GIFFORD: Deb, I think it's in the appendix. We're looking for it right now.

CO-CHAIR PARTRIDGE: Okay. Are we

ready to vote on variability? I'm sorry, validity.

MS. QUINNONEZ: Okay. Voting is now open for validity including specifications consistent with evidence, testing and threats adjust, exclusions, risk adjustment stratification, meaningful differences, comparability, multiple specifications, missing data. The options are one high, two moderate, three low, and four insufficient.

Okay. We're still waiting for two votes. All votes are in. Voting is now closed. The results read 33 percent voted high, 50 percent voted moderate, 17 percent voted low, and 0 percent voted insufficient.

CO-CHAIR PARTRIDGE: Next item or category is feasibility. We've talked about this a tiny bit. Linda was excited that it was short which is the whole point. Are there other comments? Are we ready to vote? Yes?

MS. QUINNONEZ: Voting is now open for Measure 2614 for feasibility including data

generated during care, electronic sources, and data collection can be implemented. And the options are number one high, two for moderate, three for low, and four for insufficient.

We're waiting for two votes. One more vote. All votes are in. Voting is now closed. The results are 28 percent voted high, 72 percent voted moderate, 0 percent for low, and 0 percent for insufficient.

CO-CHAIR PARTRIDGE: Finally we are at usability. Usability, again, any comments? Okay. Desi, over to you.

MS. QUINNONEZ: Voting is now open for Measure 2614 usability and use, accountability, transparency, and improvement progress demonstrated and benefits outweigh evidence of unintended negative consequences. The options are one for high, two for moderate, three for low, and four for insufficient information.

Looking for two more votes. All votes are in and voting is now closed. The results are 28 percent voted high, 61 percent voted moderate,

11 percent voted low, and 0 percent for insufficient information.

CO-CHAIR PARTRIDGE: Before we go on to the final question of recommending endorsement, Peter?

MEMBER THOMAS: Real quick, it says, there's a note on the bottom of Page 12 that says that although there's no public reporting now that ACHA is planning on public reporting with the CoreQ measures as part of the quality initiative for 2016 through '18 with 9,600 SNFs.

I'm just wondering how that compares or contrasts with the measures under the Medicare program that CMS is implementing on the IMPACT Act. Are they duplicative of one another? Are they done together? Or what's happening with that?

DR. GIFFORD: IMPACT Act does not require satisfaction. We have shared this with CMS and have asked them that they add this to five star at some point in the future. I know five star is thinking about, in the next five

years, probably adding satisfaction at some point.

MEMBER THOMAS: Thanks.

DR. GIFFORD: As an organization, we support and endorse that. Why we undertook this whole process.

CO-CHAIR PARTRIDGE: Okay. So final question is whether or not we recommend that this measure be forwarded for with recommendation for endorsement by MQF. And Desi?

MS. QUINNONEZ: Voting is now open for Measure 2614 for recommendation for overall suitability for endorsement. Option number one is yes. Option number two is no.

We're looking for two more votes. We're ready. All votes are in. The results are 94 percent voted yes and 6 percent voted no.

CO-CHAIR PARTRIDGE: In the past, as we have here, three measures that are almost identical, they're not obviously identical but almost, we have taken the votes on the first measure. And said after discussion, we first

discussed and then made a decision as to whether we would be comfortable about accepting the votes on those, on the first measure. And moving the scores, essentially saying the scores would be the same for the second and third, we don't need to go through all the individual items.

Before we even consider such a step, I want to give the people who have so patiently been working and studying as lead discussants, both Number 2615 and 2616, an opportunity to say what they want to say. And that's specifically at Lisa and Sharon Cross and Lisa Morrise. So Lisa, do you want to start? Lisa Suter will start.

MEMBER SUTER: Thank you. This is Lisa Suter. So I actually think that's a great recommendation. Because, having listened to the discussion of the first measure, I actually find 2615 which is for the long term resident measure less controversial, I think, in terms of its exclusions to the denominator than the prior measure. And the liability and validity testing

is very similar across the board.

So I actually think it's a very reasonable recommendation to translate scores from, at least my scores from that initial vote to this second measure as well as to the third measure which assesses family member care. Family member inputs, excuse me.

MEMBER MORRISE: I agree. I don't see that there's that much difference. In fact, in reading through all of them, I thought this is the same thing over and over again. It's great.

MEMBER CROSS: I agree. I was thinking the same thing as I was reading through them, that they were all very similar. So there's nothing additional that I think I'd want to add to the conversation at this point.

CO-CHAIR PARTRIDGE: Then I think we should have a formal vote that we do that. Peter?

MEMBER THOMAS: On both the long and short measures, you don't, one of the exclusions is guardians that respond, correct?

DR. GIFFORD: Yes, the legal guardians.

MEMBER THOMAS: Guardians.

DR. GIFFORD: Yes or court appointed guardians.

MEMBER THOMAS: So at first, I was a little concerned about that. Then I read that the next measure is the family.

DR. GIFFORD: Correct.

MEMBER THOMAS: And I don't remember whether guardians are included in -- I guess guardians, that wouldn't be relevant there. So it's really not related. I was thinking about someone else answering on behalf of the patient. But --

DR. GIFFORD: We don't let anybody answer on behalf of the patient.

MEMBER THOMAS: Because it's patient reported, you want to make sure that they are the ones doing it.

DR. GIFFORD: Yes.

MEMBER THOMAS: Thank you.

MEMBER SALIBA: This is Deb. The only concern I have on the long stay one was it looked like -- and I may have misread your data. But it looked like you were allowing the facility member to fill out the survey for the individual if they were unable to do it themselves.

DR. GIFFORD: Well there's no way to stop them. But if they indicate as such, their data is excluded.

MEMBER SALIBA: Okay. So I was misreading it. It looked like it was allowed. And that, obviously -- so are individuals -- how are we making sure that there is no provider influence on scoring with these? I'm assuming the short stay people have a little bit more autonomy. With the long stay people, how are you assuring that they're sort of being able to independently answer these surveys?

DR. GIFFORD: Well there's no way of assuring it Deb.

MEMBER SALIBA: Okay.

DR. GIFFORD: I mean, but it's

administered, it's in the rules and it's in the guidance that's out there. But as you know, a lot of places do provide some level of assistance out there. You know, we're trying to push that out there.

MEMBER SALIBA: So it does not require to use an independent --

DR. GIFFORD: Yes, they have to use an independent vendor.

MEMBER SALIBA: To analyze them or to administer them?

DR. GIFFORD: Well to both. But independent vendors often have the facility distribute the survey and then collect them and mail them back. So at that point, you can't really -- I mean, I can't assure one way or the other.

But yes, this requires an independent vendor to both make sure the samples being drawn and the things being calculated as being administered. But they're not coming in and doing and doing face to face or doing it in the

building.

MEMBER SALIBA: I mean, I guess that's just the biggest concern, you know, in terms of threats to validity with the long stay ones, would be someone's comfort level if it's being handed to them by a facility staff member and then collected by a facility staff member. As you know, I mean that could be a little bit of an issue.

DR. GIFFORD: I agree. It's just inherent with the whole population.

MEMBER MORRISE: I agree that it could be an issue. I also had a concern as I was reading through it and they had to dig deeper to find that it is allowed in the administration of this survey for somebody to act as a scribe which I see as being very important. Especially since my daughter has upper trunk CP and doesn't write well at all. And so, she has me fill out everything for her.

So I realize that could have some bias that enters in. Because an individual could fill

out something other than what is being reported. But that's probably minimal.

DR. GIFFORD: We do ask did someone help, yes, no? And then we ask how they helped. And so, if it's in your scenario and you're just translating the thing, yes we will count them. So we wanted to make sure that we didn't lose individuals.

MEMBER THOMAS: In this day and age of electronic everything, this is a mailed survey? Good old fashioned U.S. mail, right? I mean I don't doubt the fact that --

DR. GIFFORD: We've not tested the app for it yet.

MEMBER THOMAS: I doubt you'd get very far in this population.

DR. GIFFORD: Well it is interesting you say because a number of the different survey vendors out there that approached are using different vehicles for how to administer it. And as we're collecting data is getting much wider use, we're looking at whether the method of

collecting data differs in any way. But right now it's a paper version, yes.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: Given the population that you're talking about, is there any accommodation for patients with visual impairments or language? Will this be offered in other languages?

DR. GIFFORD: Right now it's offered in English only. Other than a handful of SNFs, would that really cross a sample size? But it definitely is something we would want to pursue down the road to explore that on how to get that broader. Especially as the demographics are changing and the cohorts are aging in other demographics. We'd like to do that.

CO-CHAIR PARTRIDGE: Sharon?

MEMBER CROSS: Regarding the family measure, I just wanted to ask for clarification about the comparison to related or competing measures. Because there was a mention about a CAHPS nursing home family. Can you just respond

to that?

DR. GIFFORD: Yes. That was before it was withdrawn for endorsement. We submitted this prior to the March 3, 2016 NQF.

MEMBER CROSS: Yes. And I'll just, in comment on that, we reached out to the AHRQ folks repeatedly. And they indicated they did not want to continue endorsement because, while they're keeping the survey and they're maintaining the survey, as Dave already mentioned, it's not being used.

We also followed up with CMS. And CMS said they had no plans to be putting that into any of their programs any time soon. So you know, that's right, we're dropping that.

CO-CHAIR PARTRIDGE: As a formal matter, Sarah tells me that we should vote, to do only the final vote. That is, do we recommend for endorsement both Number 2615 and 2616. So we'll go ahead and do that vote.

MS. QUINNONEZ: Voting is now open for the recommendation for overall suitability for

endorsement for Measure 2615. Option number one is yes --

CO-CHAIR PARTRIDGE: Before we vote, Linda?

MEMBER MELILLO: Yes. I just, if I understand correctly, you're mailing these to their homes? Yes?

DR. GIFFORD: The discharge home, it's going to whatever address the SNF has on file to communicate with them. For family members, it's going to the primary person that they have on record. And for residents, it's going to the resident who's living in the home because they're not, that is their home.

MEMBER MELILLO: Okay. So for those that have no potential to discharge, are you serving them as well?

DR. GIFFORD: On the long stay, yes.

MEMBER MELILLO: So the institution then would administer the survey?

DR. GIFFORD: You know, actually it tends to be a little bit different depending on

the survey vendor. Some will mail it to the nursing home so it arrives in their mail. Others will go around and deliver it to them in their room and let them then submit it.

MEMBER MELILLO: Okay.

CO-CHAIR PARTRIDGE: So I think now we are back to you Des.

MS. QUINNONEZ: Okay. If there are no more comments, we will have open vote now for Measure 2615 to score the recommendation for overall suitability for endorsement.

CO-CHAIR PARTRIDGE: One is yes?

MS. QUINNONEZ: We're still waiting for three votes. All votes are in. Voting is now closed. The results read 94 percent vote yes, 6 percent vote no.

CO-CHAIR PARTRIDGE: Okay. And going on --

MS. QUINNONEZ: The voting is now open for overall suitability for recommendation for endorsement for Measure 2616. Voting option number one is yes, voting option number two is

no.

Looking for two more votes. All votes are in. Voting is now closed. The results read 94 percent voted yes and 6 percent voted no.

CO-CHAIR PARTRIDGE: Thank our friends from AHCA for taking us through so expeditiously.

MR. MULLER: Thank you very much everybody.

DR. GIFFORD: Thanks Nick.

DR. CASTLE: Thank you. Take care.

(Whereupon, the above-entitled matter went off the record at 2:44 p.m. and resumed at 3:03 p.m.)

CO-CHAIR STILLE: We're going to start again, and what we're going to do -- for the six measures developed by UDSMR, is split them into two groups, and actually the agenda reflects that.

There are three that apply to skilled nursing facilities and then three more that apply to long term acute facilities. And the three in each group are change in self care score, change

in mobility score, and change in motor score, all from the FIM instrument if I remember correctly.

So similar to what we've done in the last hour, we'll have a rather extensive discussion on the first one and then have some discussion about how much that could be applied to the other two in the group.

There are many primary discussants. And so I think there will be a lot of discussion with everyone. We will start with Measure 2769 which is up on your screen right now.

(Off microphone comments.)

CO-CHAIR STILLE: Yes. Sure.

MS. THEBERGE: Do we have the audio working again, Operator? Deb, Dawn, can you hear us and can you speak?

MEMBER DOWDING: Oh hi, yes I can hear you now. Thank you.

MS. THEBERGE: Okay, great. Thank you.

MEMBER SALIBA: Yes, we're good. We're good. Thanks for checking.

MS. THEBERGE: Thank you.

CO-CHAIR STILLE: Okay. So we'll start with Measure 2769 which is the first skilled nursing facility measure. The functional change, change in self-care score by UDSMR.

The primary discussants will be Beth Averbeck, Dawn Dowding, and Peter Thomas. And we have the measure developers here, so take it away for about three minutes.

MS. DANN: Okay, thank you. I'm Kathy Dann, director of operations for UDSMR. On behalf of UDSMR and our subscribers, we are pleased to be here to offer for endorsement functional quality measures for both the skilled nursing facilities and the long term care populations.

These measures have already been endorsed by NQF for inpatient rehabilitation facilities. So endorsing them for these two populations will allow standardization of measurement, of function across all of post-acute care using a common metric which in effect

fulfills the charge of the IMPACT Act.

With me today I have our Director of Research, Dr. Paulette Niewczyk. Am I too close to the microphone? Okay, good. Who is here to answer any questions related to the measurement development process and to answer any questions related to the testing of the measures data that we provided to you already.

I have Dexanne Clohan, a doctor of physical medicine who also has an expertise in functional assessment here also. On the line we have Dr. Maggie DiVita who did some of the analysis that we might defer to for some of your questions.

Now these measures were developed out of extensive research and experience with a functional measure known as the FIM instrument. It's already mandated by CMS for rehab hospital reporting.

Here are a few of the measure's key features. First and foremost, the measures utilize a standardized set of items,

calculations, and risk adjustment methodology regardless of post-acute care setting.

This will provide the ability to compare and contrast post-acute care sites and outcomes at a common metric, a truly apples to apples comparison.

Secondly, the measures are based on data that's been collected for over 20 years in all these sites. Third, the measures require data collection of only 20 unique items that already has training materials, industry knowledge available for public use.

And lastly, while only mandated for use in the IRFs as part of the Medicare payment system already, CMS does have the rights to a royalty-free license to utilize a FIM instrument in any setting.

We thank National Quality Forum and the person and family centered care committee that is here today for their time, and for their consideration of our measures. We stand ready to answer any questions that you might have related

to our university affiliated not for profit organization, the measures we've submitted, or the FIM instrument. Thank you.

CO-CHAIR STILLE: Thank you for that very concise presentation. Thanks. I would like to have any discussion from the primary discussants.

Again, start general as much as possible, and then after everybody's sort of talked about general things, we'll go into some of the more specifics about evidence and stuff like that. So Len is first.

MEMBER PARISI: I didn't review the first measure, I'm not one of the lead discussants of the first measure. But this is an overall question about the use of the instrument. Is this going to be in the public domain or is this now something that still has to be purchased using the ability to use the FIMs instrument?

MS. DANN: The ability to use the FIM has been offered to CMS. The exact term public domain needs to be debated. The reason for us

not putting it in the public domain already is simply because we want to maintain the integrity of the instrument through uniformity.

We are offering it free, but that's the concept of the integrity of the instrument, gets maintained through uniformity.

MEMBER PARISI: So it will be free.

MS. DANN: Yes.

MEMBER PARISI: Okay. I just wanted clarity on that. Thank you.

CO-CHAIR STILLE: And Linda?

MEMBER MELILLO: The FIM as it stands in rehab requires certification of those who administer the tool. Is that something that would also be required for these?

MS. DANN: CMS does not require that certification. UDS provides that as part of its services to its subscribers. So I don't believe that CMS would require something they don't require today for IRFs in the future.

DR. NIEWCZYK: There are training manuals available for use. So certification

would be something that if you were subscribing to our services, our data reporting services would be requested. But not for use.

MEMBER MELILLO: So for the data reporting component of this, they can get that for free as well?

DR. NIEWCZYK: Not through us.

MEMBER MELILLO: But there's a venue for that?

DR. NIEWCZYK: Yes. There could be other than use for that. Correct.

MS. DANN: Just the way the IRFs now submit the data to CMS and they're going to have IRF quality reporting, if they were to choose to use the FIM, they would have the rights to report on that.

MEMBER MELILLO: Or other health information systems for instance, yes.

CO-CHAIR STILLE: Beth, you're one of the primary discussants. Why don't you go?

MEMBER AVERBECK: Thanks. And I would encourage those of you that also reviewed this to

chime in as well. The measure looks at the change in the self care on admission and on discharge. I think it's really important that we have some measure of assessment for those patients that are in skilled nursing facilities.

I did have some questions going through that are just more general questions around some of the specifications and then a couple of the exclusions and some missing data, so maybe I'll go ahead and just raise those right now.

DR. NIEWCZYK: Sure.

MEMBER AVERBECK: I'm curious as to why a 12 month measure was used for a short term stay. I mean, most of these patients would be discharged beforehand. I'm wondering about what happens for patients that are readmitted to the hospital, you know, that if they are excluded if they're under 18 or if they die, but what happens with hospital re-admissions and those patients.

And then there was a comment around if data is missing, so if we don't have complete

data, then those patients are excluded. But I wasn't able to tell based on the information submitted what percent does the same across facilities? Is it low, did it not --

(Simultaneous speaking.)

MEMBER AVERBECK: -- just curious around how the missing data is accounted for and whether that might affect the results as well. So I think those were the main questions I had as I went through this. Thank you.

DR. NIEWCZYK: Thank you. And I'll try to answer all your questions. Maggie, I may ask you to chime in on a couple of them.

In terms of the age, the measures could be used by patients of any age. However, we have this at 18 and older just because we understand that sometimes, you know, pediatric populations require different modifications to items and there's the function of development versus impairment and so forth.

So to simplify things, we kept it as adults early. Clearly, if a patient dies while

under care, we don't have any discharge scores to go on. So we have an admission but we have nothing else so certainly can't assess what the change, you know, would have been.

And that's a typical standard way that others have within the industry have excluded on that basis. In terms of missing data, we actually have a very, very small percent. Maggie, do you happen to know on average how many cases each year would be missing? I would guess about less than two percent, but I don't want to make something up here.

MS. DIVITA: In terms of missing the required elements for the measure, I would say zero because they're required in our system. So something like creating age from birthdate, we need that information, or all of the items are required items. And so there really isn't missing data as it stands.

DR. NIEWCZYK: So they could refuse to do the assessment altogether on a patient but they can't, like, --

(Simultaneous speaking.)

MS. DIVITA: They could but then, yes, but then it wouldn't be transmitted through our system. So there would be really no way for us to understand or to know, you know, the percentage of population they decide not to collect the data on.

My gut would be that it would be very small, because otherwise I'm not sure why they would bother, you know, utilizing the measure essentially.

DR. NIEWCZYK: And Maggie, can you also speak to the 12 month window?

MS. DIVITA: Yes, I am, it's a good question. It's been quite a while now since we developed these. You know, I think that we tend at UDS to look at things in a really 12 month way to sort of look at how people are doing.

We wanted to give them a long enough timeframe to allow for, you know, a good number of patients so that you're not looking at a smaller number of patients. And so I think it's

just sort of something that we tend to utilize at UDF.

And so the 12 month timeframe sounded pretty good to us. It also gives, you know, a good amount of time and are we talking about SNFs right now or are we talking about long term care? I can't remember. But in general --

(Simultaneous speaking.)

DR. NIEWCZYK: We're talking about SNFs, but it's the same method for both.

MS. DIVITA: Right, yes. It gives us a good amount of time to follow through. You know, average length of stays aren't nearly a year long, but that 12 month period should allow us to get a good view of a true length of stay.

DR. NIEWCZYK: And certainly internally we do look at things by quarter as well. Our facilities do get quarterly reports. But as we indicated in the submission files, facilities had to have at least 30 cases over a year, over that 12 month period.

So it's really a function to allow

those smaller facilities to have an adequate case count for representation.

MS. DIVITA: Correct.

CO-CHAIR STILLE: Dawn, you were one of the primary reviewers as well. Would you like to chime in?

MEMBER DOWDING: Oh yes. Hi. My main question was to do with issues to do with data burden because a lot of the measures, the things that you're collecting on the FIM are actually included in the MDS assessment and would also be in the catch all when it institutes this hostile approach to acute care settings.

And I guess my main concern with all of these measures is that we're, you know, is it going to be something that is instead of the CARE tool or replace those item measures because otherwise there's increased data burden on the users in terms of having to complete these measures more than once.

DR. NIEWCZYK: So I can try to speak to that. Yes, you know, certainly we understand

that function as something that's assessed in other instruments and often a variety of different ways.

Our measures that we have submitted here today are part of a larger tool which is the FIM. Certainly, you know, when you get through the measure files, you'll see under concurrent validity that was our gold standard. Right?

So how did those measures compare to the full FIM. But this is not, you know, it's not identical to the FIM. How it relates to those other instruments, be it MDS or to the CARE tools, those are measuring components of function as well, although some of the items are quite different.

You know, for instance, I'm not sure what measure we're talking, I think it's self-care. But for self-care, we have both cognitive items as well as physical function, physically function items.

And the CARE measure for self-care includes only physical functioning. There's no

cognitive items. In terms of the MDS, you know, the MedPAC reports from 2013 stated that skilled nursing facilities did not demonstrate any change over 2011 to 2013 from those functional items.

So we certainly, and as is indicated within our files, we did see a great deal of change in terms of some of the functional values over the years that have been indicated. So I believe they probably measure different functional domains.

CO-CHAIR STILLE: Peter?

MEMBER THOMAS: I've got a -- I'm sorry, was there a follow up?

MEMBER DOWDING: Yes, I mean, I'm not disputing that they might measure different things. But a lot of the items of information that go into the different tools are the same.

So are you going to ask them to collect the same information twice, or are you going to use the information you collect for the FIM or for CARE tools to feed into the methods because otherwise you're asking a group of very

busy clinicians to fill out more than one item to do somebody's function is my concern, that's all.

MEMBER THOMAS: So the last set of measures excluded readmissions to either an acute care hospital or being sent over to a different post-acute care setting as well as death of a patient. How do you handle, in this measure how do you handle those readmissions or transfers to another setting of care? Are they excluded from the calculation or not?

DR. NIEWCZYK: I do not believe that that's an exclusion criteria. Just those that are under ages 18 and who had died while in care.

MEMBER THOMAS: Right, those are the only two?

DR. NIEWCZYK: That's the only two.

MEMBER THOMAS: So I'm just wondering, I mean, which I'm in favor, I appreciate that. I'm just wondering how would you score if someone gets readmitted to the acute?

DR. NIEWCZYK: Oh, excellent. Yes, we do have one of the additional data collection

points that's not part of the measure is discharge location.

So clearly, we do ask about discharge back to the community, discharge to another post-acute care venue like maybe inpatient rehab or home health outpatient care, back to acute care. So all of that is still collected as part of data elements.

MEMBER THOMAS: Okay. So last time we were here and looked at a whole other set of measures, I raised this same issue, I'm going to raise it again now and that's the Jimmo versus Sebelius settlement.

And that is a case that states that Medicare law is, and CMS agrees that Medicare law is or requires that skilled nursing facilities, outpatient therapy, and home healthcare provide services, skilled care, skilled therapy or nursing, to an individual, beneficiary who may not be improving in order to maintain their function or prevent deterioration of function.

And I'm less concerned in the IRF

because that settlement really doesn't do all that much in the IRF setting. But I'm very concerned in the SNF setting because it does apply to SNFs.

And if you've got a quality measure that's based on, and you say it a number of times in your document, the emphasis on restoration. Now you do admit, I will admit you say or maintenance of function which is important. But in other areas you say the primary benefit of rehab is functional improvement, improvement, improvement.

And I'm hopeful, I'm trying to figure out a way that you don't create a quality measurement tool that is going to disadvantage individuals who won't necessarily improve but still gain something from their experience in a SNF.

And those folks are going to become persona non grata because that SNF doesn't want to have those people in its data mix. Can you tell me how you address that fairly difficult,

what I consider to be a fairly difficult problem?

DR. NIEWCZYK: So that's an excellent question. And our measures are sensitive to the fact that patients may gain, may not change, so maintenance, or may decrease in function.

So certainly an aggregate score would be at the measure level, right? But you can also look at each item level. So you may see that a patient improved in one item but decreased in another. If you look only at that measure score, it's going to show no change, but there was some changes.

So it works at both the patient level as well as the population or facility level. Additionally, it's understood that it's a change, it's not a gain. So these measures are not functional gain. These are functional changes. So it could be in either direction.

We understand that in skilled nursing in particular, you have sometimes two patient populations where you have those that are there for restorative care and those that are there as

residents that might also be getting some rehabilitation.

Within those residents, in all likelihood, you may be seeing more maintenance of functioning, which might be the goal, just not to decrease. So certainly our measures can work in that capacity.

And facilities know which patients are which. So there would be, you know, no difficulty in building in an additional variable asking is this patient here for restorative treatment or patient or resident? And certainly we could then look and stratify on those basis.

MEMBER THOMAS: I don't know if this is appropriate to ask, but in the documents, in the writeups as you proceed there's a number of measures that have the same concern that I've got, that has been raised.

Can you please, can we encourage or require there to be some recognition that it's not just about improvement in function and that, you know, these tools should not be used in any

way to bias admission against people who won't improve --

DR. NIEWCZYK: Absolutely.

MEMBER THOMAS: -- something to protect against that.

CO-CHAIR STILLE: Yes, yes. I had the same concern with 2775 which it says change in motor function, so same kind of thing. Whether it's stratifying among people who are expected or not expected or something.

DR. NIEWCZYK: So certainly we can add some language in, you know, in the documentation stating that the measures could be used, you know, to assess maintenance of function as well as to identify decreases in functioning.

We have users in some outpatient as well as adult day programs. And really the goal there is maintenance. But they sometimes see patient populations that will decrease function. And then it's an early warning sign to add additional patient referrals or resources, or maybe to modify the patient's care plan.

CO-CHAIR STILLE: Could I ask one more real quick? And that is my understanding, and correct me if I'm wrong, is that the FIM really has a very strong connection to the inpatient rehab hospital and unit.

It was created around that patient population and that setting, and it was really geared toward trying to measure function in that setting. So is it really that easy to just take that same functional measurement tool and apply it to a SNF or an LTAC? It's an open question. I'm not doubting that it is. But please explain how --

DR. NIEWCZYK: In post-acute care it is. But what we found through the 20 plus years, I don't know what's happening to my mic, but the 20 plus years of study on the tool is that function can be measured and can be assessed with a very small number of items.

Now it's not all-encompassing, it's not a perfect instrument in any capacity. But a rough estimate in terms of where the patient's

functional status is, and the fact that you could use it in other post-acute care venues, even those like the long term care facilities where you may not see drastic improvements in functioning.

It gives you a point as to where that patient's status was when they came in, when they left, and then when they went to another, you know, downstream post-acute care venue using the same items with the same rating scale.

CO-CHAIR STILLE: Thank you.

MS. DANN: Could I just add to that that with our submission, I believe we put a bibliography of something called burden of care, and the instrument can measure that latent trait in any venue.

Therefore it is viable to be used in any venue because it can measure the burden of care regardless. And I don't know how many publications are in that bibliography. I think it's 15.

DR. NIEWCZYK: I think we gave a very,

we give an annotated one, yes.

CO-CHAIR STILLE: Beth?

MEMBER AVERBECK: Just a follow up on a question to Peter's comment on the patients that aren't improving. Does a case mix adjustment then account for that, that what you might expect since it's a measure to expectation that if someone wasn't seeing improvement and a change that would be taken care of by the way that it's case mix grouped?

DR. NIEWCZYK: So certainly case mix may come into play, impairment group code would likely be a greater indicator. But, you know, oftentimes it's really hard. Those are things that are probably more obvious in the medical records. But yes.

CO-CHAIR STILLE: Okay. Any more general comments before, oh I'm sorry, Linda?

MEMBER MELILLO: I'll come back to it.

CO-CHAIR STILLE: Okay. And yes?

DR. CLOHAN: I just wanted to make a comment about Mr. Thomas' remark which is very

pertinent on the Jimmo versus Sebelius case. And I think it's a very good suggestion to put something in NQF's write up as well could enter that.

The fundamental question you raise though is endemic across all quality measures of whether there could be an inadvertent consequence of a quality measure of suddenly making a patient who is at higher risk to "fail" that quality measure become less desirable to a provider.

You can look at it on new or worsened pressure sores, you can look at it at fall risk, you can look at it on do you want to take somebody in your site who has refused a pneumonia or flu vaccine.

It's an inherent risk. But I think there will be good protection on this one, and I think it's important for us to keep separate in our minds the coverage criteria for Medicare or other payers that entitle people to coverage and care.

And that's what Jimmo sort of

addresses, that people are entitled to that level of care whether they're expected to improve in their function or whether the purpose of that care is to maintain or even slow deterioration. It could be that a person is known to, that their function will deteriorate but medical intervention could slow that deterioration.

So I think it's great that you brought it up, and I think this organization, everybody else interested in quality measures always has to be mindful of potential for unintended consequences. And this one I think you can help by having, making sure your literature addresses that point.

MEMBER MELILLO: So one of my questions is I have a fear that this data may be used to try to compare the various levels of post-acute care. And they are very different in terms of what a patient or family could expect in terms of services and what patients and families can expect in terms of length of stay.

I mean, all of these metrics vary

substantially. So that is a concern of mine. And whether or not the patient had an interim stay in an IRF before they went to a SNF. These levels of care are not interchangeable is the point that I'm trying to make. And so I'm just -- be cautious of that.

DR. NIEWCZYK: We sort of believe that too. But unfortunately, there's many others that don't believe that, that think that post-acute care is maybe the same across the board.

I don't think we'll ever really know until we collect it, collect outcomes in a standardized way. But it needs to be the same types of questions asked upon the same patients in all venues.

It can't be this item applies to this venue but not this venue. Otherwise you won't know, right?

MEMBER MELILLO: Yes, there is some research out there, some new research that I'm sure you're familiar with that helps to differentiate. But I hear what you're saying.

But it is a concern of mine that if we have these like measures, that they're going to assume that they're measuring the same thing. And they're not but they are. You know what I mean?

DR. NIEWCZYK: Well, function is important I believe to all persons.

MEMBER MELILLO: Of course.

DR. NIEWCZYK: Whether you're a patient in a facility or, you know, you're sitting here like us today. So I believe it's important in, you know, all capacities.

So looking at some aspects, some very basic aspects of function in terms of things like communication or expression and, you know, ability to eat and how you're, are you eating with assistance of a tube or not. All of those factors do mean something.

And certainly, they can be asked in all of those venues because they deal with patients that have those issues in all of those venues.

MEMBER MELILLO: So I don't know if

this is the right time to have this discussion. But I did want to also mention the data that you used, the data sets.

DR. NIEWCZYK: So these should be really general.

MEMBER MELILLO: Okay.

(Simultaneous speaking.)

CO-CHAIR STILLE: Yes, we'll talk about data in probably section two. Yes, okay.

MEMBER THOMAS: Really quickly, I just would quarrel somewhat with the fact that your statement that we don't really know. We know what's required of an IRF.

DR. NIEWCZYK: Oh yes, yes.

MEMBER THOMAS: We know what's required of a SNF.

DR. NIEWCZYK: The care, yes.

MEMBER THOMAS: We know what's required. So there are different levels and intensities and coordination of care being provided now. We all know that.

We may not know how risk adjusted, how

they all correlate based on outcomes and I get that. But the fact is these are different products.

DR. NIEWCZYK: So there's parameters what's differentiated both from a SNF, right? But certainly, like, a hip replacement could be treated at either venue. What are the outcomes? Did one perform better? Was the care the exact same?

We don't know. The cost certainly is different, right, and the intensity of the rehabilitative services could be different. Some skilled nursing facilities believe they're just at par with the requirements of an inpatient.

The point is we don't know what those outcomes truly are. And if we're looking at measuring function, I think we need the same types of items within those settings to be able to assess that.

CO-CHAIR STILLE: Sam, go ahead.

MEMBER BIERNER: I just wanted to say that it's not just, I fully understand what you

said about the function. But when we're looking at different levels of care, there are other issues such as the medical complexity, the acuity of the medical problems that the patient has and so forth that are not always captured directly in this instrument which was intended just to look at the patient's physical and cognitive function. I understand that.

But I just wanted to make that clear in case there are others who aren't as familiar with the difference in the IRF versus the skilled nursing and other things.

DR. CLOHAN: And thank you for that clarification because it's very true. I share your concern of sort of the potential for wrong use of any measure.

But I think probably as a Committee and as NQF and as all of us operating under the IMPACT Act, pretty much the marching orders have been put out there for site neutral measurement techniques.

And so I think we feel very proud to

be bringing you measures that have been used in all of these different settings. They've been used in, you know, we're talking specifically about one use in skilled nursing facility now. But you'll see all these six measures, three different functions being measured and then two different sites over which they're being measured, proposed to be measured.

And so the instrument has worked effectively in all of those settings, then I guess it becomes the next stage after endorsement to make sure that these are used properly.

And if some day somebody says oh, let's compare this with this, we know the measurements are equal, apples to apples. What conclusions people may draw from that if they don't properly look at different populations, different intensity of care, they could be drawing wrong conclusions.

But I think the IMPACT Act pretty much sets the requirement out there for us to find common measures. And we believe this is one

that's very strong, has a long history behind it and a lot of data, published research.

And these measures can actually be used, I'm very proud of them because they can be used for the purpose that I think we all believe quality measures are for, not just a measurement but giving us some leverage to improve quality in the future.

So in 2015 there was an article published using FIM data and comparing that with patients who were found to have new or worsened pressure ulcers. And it was discovered that these functional measures correlated with predictive quality.

And so that's what I think the quality measurement game is all about, to tell us something that is then actionable. So then we can say oh, we've identified some people who are particularly vulnerable. So that's what we're hoping this can do.

CO-CHAIR STILLE: Okay, thank you. I think we're ready to move on to start talking

about the evidence. So let's start with 1A, evidence for the measure. Any specific discussion that hasn't already been covered?

MEMBER THOMAS: I already said everything, all the comments I had.

CO-CHAIR STILLE: They're in there? Okay, we're good. Okay. Should we vote on evidence then, if I can find my clicker. There it is. 2769 1A evidence.

MS. QUINNONEZ: Okay, we will now be opening voting on Measure 2769, functional change in self care score as for skilled nursing facilities, and this is for the evidence. Rationale supports the relationship of health outcome or PRO to at least one healthcare structure process, intervention, or service.

Option number one is yes, option number two is no. Looking for two more votes. Okay, all votes are in, voting is now closed. The results of this vote was 100 percent yes, zero percent no.

CO-CHAIR STILLE: All right. And we

will move on to performance gap. I actually had a question. And mine as primary reviewer is in a couple more, but it's still one of the SNF ones.

You mention that almost half of facilities are below expectations in 2014. But I can't find what expectations are.

DR. NIEWCZYK: Okay. So what we did is we took all of the facilities and I believe we used 174 facilities, skilled nursing facilities. And we took the admission and the discharge score within the Rasch modeling.

We put them together so it, a progression to the mean in essence, to come up with where an average patient's function would be on each item within that measure.

So that gave us our, like, expected values. And then you could look to see a given facility, were they above that threshold or below? And then were they that way at admission and then again at discharge.

CO-CHAIR STILLE: Okay.

DR. NIEWCZYK: So that way we could

see if patients had been maybe low before and now they came up really high. Certainly they would have --

CO-CHAIR STILLE: So it's not a benchmark but rather basically a statistical place where all the other places are. Okay.

DR. NIEWCZYK: Exactly.

CO-CHAIR STILLE: Okay, great. Thanks.

MEMBER MELILLO: Did you include IRF and LTAC into that?

DR. NIEWCZYK: Not into the, so I know in the measure files you have it broken down so you're able to see what the values are for IRFs as well as LTACs. But we had those separate.

MEMBER MELILLO: Okay.

CO-CHAIR STILLE: Others? Lisa?

MEMBER SUTER: I may have missed it, I was looking at some of these in more detail than others. But anywhere do you provide actual distributions of the facility level scores? I know you provide the average and you provide sort

of percent above and below expected.

But it's a little bit hard to tell what the actual distribution of scores at the facility level is and do you have that information anywhere in any of the applications?

DR. NIEWCZYK: I don't believe we included it in the measure files, but certainly we're able to do that internally, yes.

MEMBER SUTER: I think that would be really helpful because in determining a performance gap, I will tell you I really can't interpret the information that's provided.

CO-CHAIR STILLE: Right, I agree with that. You know, if half of the facilities are below expectation but it's just below the mean or whatever, okay what's the distribution. That would be super helpful. Sherrie?

MEMBER KAPLAN: Yes. Lots has been going on over the last six years to try and improve functional status. Do you have any data on whether or not this change in function is mutating in response to efforts to improve

quality?

DR. NIEWCZYK: We have noticed changes over the years, yes, not only in terms of rating and scoring patterns but also in terms of patient populations.

MEMBER KAPLAN: But your data don't show it. Your data look rock solid stable.

DR. NIEWCZYK: If you stratify certain ways it becomes pretty clear. It's just that, you know, we've already incorporated a ton of information already, so we didn't want to belabor you with so much more. But yes, I mean, certainly we could provide you more specific if you want.

CO-CHAIR STILLE: Any other discussion of performance gap? Okay, should we vote?

MS. QUINNONEZ: Voting is now open for Measure 2769 for importance to measure and report performance gaps. Data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups.

Option number one is high, option number two is moderate, option number three is low, and option number four is insufficient. We're looking for one more vote. All votes are in, voting is closed.

The results are 16 percent voted high, 68 percent voted moderate, 5 percent voted low, and 11 percent voted insufficient.

CO-CHAIR STILLE: Okay. So we move on to reliability. The psychometrics discussion. I believe that there were some questions related to this that resulted in data being submitted over the last couple of weeks about inter-class correlations. So I'll call your attention to the one pager that got submitted about two weeks ago that's in your bunch of documents.

(Off microphone comments.)

CO-CHAIR STILLE: No, no. it was available, just relatively late. It was, I think, on May 24th. Yes, it should be in there if you have everything printed out.

CO-CHAIR PARTRIDGE: Maybe you would

like to -- maybe we should --

(Simultaneous speaking.)

CO-CHAIR STILLE: Maybe we should put that on screen, or the developers can talk about it too.

DR. NIEWCZYK: I could certainly speak to it if you would like. It might be easier for me to just summarize it for you.

CO-CHAIR STILLE: Sure, okay.

DR. NIEWCZYK: So for, there's actually a couple different measures for reliability. So what's included in the specification was done on 174 skilled nursing facilities with 3,984 patients.

And this is a sample, a random sample of patients. And it was pulled on the patient level data which is why you're seeing such a, you know, variety of facilities represented.

The Cronbach's alpha was 0.92. The intra-item correlations were 0.70 to 0.84. And then subsequently we were requested by the Committee to perform inter-class correlations at

the facility level.

So we took a random sample of 25 skilled nursing facilities. They had to have at last a 30 case patient count in order to be included. And we split, each facility we split randomly into two data sets, and then we did averages at the facility level to run the ICC and compare across facilities.

So for the self-care, the ICC was, and this is between facilities, -0.03 with a P value of 0.59. So that's pretty bad. But that's actually good because it shows that there is a good amount of variability between facilities, thus we could identify, you know, performance gaps for instance.

We did the same exercise using the exact same data, and we ran within facility inter-class correlations. So within facility, the self-care ICC was 0.87 with a P value of less than 0.001 which means at the facility level there's a very strong degree of consistency in patient function ratings.

But between facilities, it varies. And that's what we see pretty much within the field as well.

CO-CHAIR STILLE: And Sherrie?

MEMBER KAPLAN: I am concerned about what you did because what you're looking for with between facility and class correlation coefficients is you want to see a thumbprint across patients within a facility and you want to see a lot of consistency within the facility and you want to see big between facility differences.

You're not seeing that. Now what you did within the facility says you have a lot of within facility, you have a lot of within patient across items consistency. So there's tight variations that are not a lot of within patient variation.

But there is a lot of between patient variation within the facility. That would drive those results. That's what your bottom paragraph says. So within a facility there's a lot of between patient variation.

But between facilities, there isn't. And there is a lot of within facility variation across patients and that's a problem for these kinds of analyses.

Another thing you could do in this kind of circumstance is do generalized estimation equations which give you approximately the kind of same between facility variation.

But the concern here is that you don't have a real strong thumbprint across patients within a facility, and that's trouble for these kinds of your ability to then make sense of any between facility variability that you see.

DR. NIEWCZYK: So the within facility ICC was 0.87.

(Off microphone comments.)

MEMBER KAPLAN: That's across patients. The within facility is across items within patients. When you're looking at a within facility, all you're seeing is between patients.

DR. NIEWCZYK: Yes.

MEMBER KAPLAN: And then across items

within a patient.

DR. NIEWCZYK: Yes.

MEMBER KAPLAN: So the error variance is within patient variability. And that's what you would expect to see given the Cronbach's alpha as you know.

DR. NIEWCZYK: Yes.

MEMBER KAPLAN: But between facility variation --

DR. NIEWCZYK: -0.03.

MEMBER KAPLAN: Yes, that's trouble.

DR. NIEWCZYK: Why would that be troubling? We wouldn't want to see a really high --

MEMBER KAPLAN: Oh yes you do, yes you do. It's exactly the opposite. You want to see a lot of tight, you don't want to see, you want to see a thumbprint across patients within a facility.

DR. NIEWCZYK: Yes, which we see.

MEMBER KAPLAN: And then -- no. And then you want to see a lot of between facility

variability. See, the coefficient is between --

DR. NIEWCZYK: I think we're saying the same thing.

MEMBER KAPLAN: No we're not. Here's the dilemma. Within facility variation looks at you got a patient level. You got patients saying there's pretty much the same thing across these items in the measure.

And then you got a lot of between patient, within facility variation. So patients, within a patient you're saying the same thing. But between patients within a facility, there's a lot of difference.

That's a problem at the next level because if there's a lot of within facility variability at the patient level, you're going to have trouble seeing between facility variability.

DR. NIEWCZYK: So it's kind of, I think you're --

(Simultaneous speaking.)

DR. NIEWCZYK: So Maggie, maybe you could speak to this better because I don't think

I'm communicating it the right way.

MS. DIVITA: So I will start off by saying that we did the analysis that was suggested to us by someone in the Committee. The ICC, just my general knowledge of it, I didn't really understand how to do it with our data which is why we had presented it with previous measures.

And so someone had suggested well, what you would do is you would, you know, randomly assign half the patients to one data set, randomly assign the other half to look at the averages across your measure.

And then do an ICC across facilities in that sense, and that's what we did. And so when we looked across facilities comparing the measures across, so between facilities, that's when we got that very low ICC.

When I changed the data just slightly so that we were comparing within, again at the measure level, not at the person level, this is average facility, you know, two data sets, two

data points essentially per facility, one random, one random and half, you know, done half, that's where we get the 0.8.

So perhaps that we're misunderstanding what we were required to do. But our, what we thought we found is exactly what Paulette was saying. So I guess I'm not following you either.

MEMBER KAPLAN: That may be a problem that NQF has in how it's communicating exactly what's wanted here because it wouldn't be fair to hold the measures developer to a standard that they didn't understand.

MS. DIVITA: Right. We just argued that the ICC, because, you know, in my understanding the ICC is looking across perhaps items or perhaps a person within Ravers, but that's not really the data that we have.

So we sort of did the ICC as suggested to us and those are the results that you see.

MEMBER KAPLAN: So Sarah, is this something that we should just then kind of ignore because it isn't clear that the information got

communicated correctly, or could we ask them to go back and communicate and re-do? How would you want to handle this?

MS. SAMPSEL: Well, I think we should do a couple things. There's also another document, Maggie, if you remember that you submitted and I think you resubmitted --

MS. DIVITA: Right.

MS. SAMPSEL: -- for this as well.

MS. DIVITA: Yes.

MS. SAMPSEL: And I don't know if that will get to some of these as well. And so, you know, Sherrie, I think this goes back to our conversation we had last Friday is really trying to understand that within and between and how we're communicating for the developers.

And you're right, we can't change the rules now. At the same time, you know, I think the Committee needs to make sure that they're comfortable with what this information is, what the data is, and is it truly, is it a good indication of quality.

And so I think, is this, no that's not it. There's another document, sorry. Suzanne, it was the one that was submitted after Phase 2 and then they resubmitted it --

MS. DIVITA: Yes, there was, it should have a bunch of graphs, right?

MS. SAMPSEL: Yes. And Maggie, you resubmitted it for this one as well I think.

MS. DIVITA: Yes. I did, yes.

MS. SAMPSEL: Yes, I think it's all the way at the end. It might be --

MS. DIVITA: So for the item level analysis across facilities to show the consistency.

MS. SAMPSEL: Page 308, 308, 309. Yes and it still might not be, Sherrie, what you're looking for.

MEMBER KAPLAN: Yes, no.

MS. SAMPSEL: But yes, I mean, so I think though that, because I would also say that this is very much the same type of information that was supplied in Phase 2 for the IRF

measures.

And as a recollection, you know, the measures were recommended for endorsement at the in-person meeting and then the additional data was still supplied, and additional questions did not come up.

So I mean, I think Sherrie, back to your question, it's kind of a consistency as well. It's a consistency in determinations and the level of information we had in front of us.

MEMBER KAPLAN: So help us then. What are we to do with the, because --

(Simultaneous speaking.)

MEMBER KAPLAN: Yes, that's not going to help, because of the way the analysis was done, this is not going to help us. So either we say you know, this isn't, there was some not clear, inclarity, unclarity in the way that the kinds of data that were being looked for were communicated, in which case we pretty much have to deal with not the facility level reliability or how do you want us to think about this.

MS. SAMPSEL: So first I'm going to ask Paulette and Maggie. I mean, from what Sherrie explained that she would like to see, how quickly could something like that be turned around?

DR. NIEWCZYK: I mean, we would be happy to get you what you want. And we, you know, we have access to a lot of data. It's just the matter that first of all, our repositories are set up at the patient level.

We know facility level information, but it's not set up by facilities. So we have, you know, for instance we could have a year's worth of data and we have a half a million or a quarter of a million cases.

They may then be from, you know, 200 facilities or 800 facilities. So then first we have to restructure our data, and then it depends on, you know, which is why we did the subset, why we pulled a sample just because we had to reconfigure everything.

We would be happy to do what you want

to do. It just depends on what it is, and how long it takes will depend on what the request is.

MS. SAMPSEL: So I mean, I think where I'm trying to go with this is kind of, you know, obviously there have been past discussions based on this level of data if not less on endorsement of UDSMR measures.

And you know, we also have, although they were less so we're not going to deal with that yet. But we would only ask for a subset of data again. So what I'm asking for, you know, so Sherrie, we know you've expressed some discomfort and not only the interpretation of this is saying that the measure is reliable.

And that really is the charge, is the measure reliable. And if you don't feel like you can make that decision, then you know, that's kind of how you have to vote on this. And we would just have to work with UDSMR to get the right data.

MEMBER KAPLAN: Well it kind of depends. Is the measure reliable at the patient

level? Looks to me. I mean, you know, we've had a lot of experience with these kind of measures at the patient level. So at the patient level it is reliable.

Then the question was at the unit being compared is it reliable. And that's where I get uncomfortable given what's gone on. And I don't know how to deal with that issue yet.

But you know, you have a lot of data. And even you said there were 20 cases per, 30 cases per facility.

MS. SAMPSEL: Thirty cases per facility.

MEMBER KAPLAN: So even on that set of data, never mind splitting it up into halves, you could do generalized estimation equations on that, you could to the intra class correlation coefficients on that.

You could do the analysis that would probably address this issue. What the results show is a whole other issue in how you interpret them. But the analysis could be done fairly

straight away I think.

CO-CHAIR STILLE: Good, so I think all the issues are laid out on the table. I think people's understanding is the best it can be. Should we go ahead and vote? Oh, I'm sorry. Linda, one more thing. Yes?

MEMBER MELILLO: Do you have data within facilities -- I'm sorry. Do you have the data within facilities stratified by diagnostic group? I know the FIM has proven to be extremely sensitive to diagnostic group.

DR. NIEWCZYK: Yes. So certainly we can look within facilities at impairment group, as well as CMG. Yes. And we do have a cross over CMG that's applied to a skilled nursing facility as well.

MEMBER MELILLO: Okay, because that might be able to break out that inter facility different spread, variability.

CO-CHAIR STILLE: Okay.

MS. QUINNONEZ: Are we ready to vote?

CO-CHAIR STILLE: I think we're ready

to vote.

MS. QUINNONEZ: Okay, voting is now open for the reliability of Measure 2769 including precise specifications in testing. Voting option number one is high, option number two is moderate, option number three is low, and option number four is insufficient.

Looking for one more vote. All votes are in and voting is now closed. The results are 16 percent voted high, 47 percent voted moderate, 11 percent voted low, and 26 percent voted insufficient.

CO-CHAIR STILLE: So it passes. Okay. On to validity. Specific points about validity, any of the primary discussants.

MEMBER MORRISE: I think my question I had earlier was around missing data elements.

CO-CHAIR STILLE: Yes.

MEMBER MORRISE: And I think that that's been answered. So thank you.

CO-CHAIR STILLE: Yes. And I'll put in my usual plug for patients under 18 in the

future. If this can be adapted to or analyzed with same, similar data for somewhat younger age groups it would be very useful.

MS. SAMPSEL: Well, and I think this isn't part of this review. But I can throw it out there.

DR. NIEWCZYK: I'll let you comment on that.

MS. SAMPSEL: We do have a separate measure called the WeeFIM, and that's meant for the pediatric population. Maybe we'll submit those measures someday.

CO-CHAIR STILLE: Yes, maybe not the range of facilities you've got here. But you know, one of the rehab facilities, sure.

MEMBER SUTER: Do you have data on the response rate at each facility? I know the reliability and validity testing was done on large numbers. But I was just wondering in terms of data collection what the average response rate is across facilities.

DR. NIEWCZYK: So that's a good

question. We don't have it because right now, the skilled nursing facilities that are electing to use our measures are doing so voluntarily.

So they're coming to us, they're using the measures on top of the MDS and any other functional measures that they or other medical indicators and so forth. So it's likely that there's an intrinsic drive that they're interested in assessing function and quality and so forth.

So they're probably capturing the majority, but we cannot be certain of that.

MEMBER SUTER: And you don't ask them to report --

DR. NIEWCZYK: We ask them to --

MEMBER SUTER: You don't ask them to --

(Simultaneous speaking.)

DR. NIEWCZYK: -- assess all patients. That's sort of what we request.

MEMBER SUTER: But to include a question that says how many of your --

DR. NIEWCZYK: Patients did you submit?

MEMBER SUTER: -- of eligible patients did you submit data on.

DR. NIEWCZYK: Yes, no.

CO-CHAIR STILLE: Any other discussion about validity? Should we vote?

MS. QUINNONEZ: Okay. Well voting is now open for Measure 2769 on scientific acceptability of measure properties and validity including specifications consistent with evidence, testing, exclusions, risk adjustment, stratifications, meaningful differences, comparability in multiple specifications, and missing data.

Option number one is high, option number two is moderate, option number three is low, and option number four is insufficient.

All votes are in and voting is now closed.

The results are 21 percent voted high, 68 percent voted moderate, 5 percent voted low, and 5

percent voted insufficient.

CO-CHAIR STILLE: Okay, we discuss feasibility. Lee?

CO-CHAIR PARTRIDGE: I was actually looking at another one of the SNF measures, but I think the issue is equally relevant here. And that has to do with -- I'm sorry. It's on the next one, usability.

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: But I'm not quite sure, yes. I think it's more appropriate for usability so I'll hold.

CO-CHAIR STILLE: Oh, we'll have a little bit of discussion there, great. Any feasibility comments? All right, let's vote on feasibility.

MS. QUINNONEZ: Voting is now open for Measure 2769 on feasibility including data generated during care, electronic sources, and data collection can be implemented.

Vote option number one is high, voting option number two is moderate, voting option

number three is low, and option number four is insufficient. Waiting for two votes.

All votes are in and the voting is now closed. Results for feasibility are 26 percent voted high, 58 percent voted moderate, 16 percent voted low, and 0 percent voted for insufficient.

CO-CHAIR STILLE: Okay, now usability. Lee, it's all yours.

CO-CHAIR PARTRIDGE: Okay. We've talked about this before in the context of the USMDR measures. But Linda raised it again. One of the important characteristics is having the staff trained in knowing how to use the tool.

And I know that for your purposes, you actually require that the staff be certified. I know CMS does not. But I just wanted to -- I don't think this, it's kind of a measure of down the road what we might expect would happen with this measure being used more widely.

It is a burden for the facility that hasn't trained their staff. And if they want to participate, they're going to have to engage in

the training. How extensive you expect that would be I don't know.

But I remember Sam raising it last time around these measures. So maybe Sam, you could comment a tiny bit too.

MEMBER BIERNER: I had put that in my written comments about review of this. The implementation in the inpatient rehab setting requires some involvement of all your therapy staff, nursing, and physician staff in learning how to do the rating and establishing good inter rater reliability.

So the groups that will be doing this will probably be nurses, patient care technicians, and therapists in the SNF or LTAC settings. So it will require an investment by the facility to train their staff.

It does require periodic retraining depending upon your turnover of staff at your facility which I think, just anecdotally I know it's somewhat higher in SNFs and LTACs than it is in inpatient units.

CO-CHAIR PARTRIDGE: I don't think that it's particularly an issue from the technical sense. It's just that I would caution that maybe other than your members and the few others, it may not be as adopted as widely as we would like.

DR. NIEWCZYK: So here's the thing. There are extensive educational materials, so there is a training guide. Certainly, you know, clinicians can self study. We have a master exam.

But again, that could, you could certainly use the tool without passing the master exam. However, what we maintain is that if you, you know, function isn't always easy to assess.

So if you want to have an accurate measure that you want to invest the time in to know about how to assess the construct of interest. It is an interdisciplinary tool.

So it doesn't matter, it's sort of discipline free. It could be done by nurses, it could be done by therapists, it could be done,

you know, by physicians.

Often whoever does the rating doesn't have to do the full measure. So you could have a couple of items be rated by one clinician and a couple by somebody else.

However, if you don't understand what you're measuring, then it's, I use the term garbage in, garbage out. So if you want to measure quality, you want to know that you're measuring function correct.

CO-CHAIR STILLE: And just to reiterate, oh I'm sorry. I skipped the line. Sorry, Nicole first and then I think Linda, Lisa, and Becky. I think. Nicole's first, I know that.

MEMBER FRIEDMAN: Sorry. One of the comments was potentially this leading to longer length of stays to increase the ability for patients to get a higher score. So of course that leaves increased cost. Is that something that has been a concern?

DR. NIEWCZYK: Well, what we've found

anecdotally is that the trend sort of follows the policy. So if length of stay is shortened than actual length of stay will be shorter.

It seems to be less driven by the patient's actual function and more driven in terms of payment. But that's, again that's out of our realm.

MEMBER MELILLO: It concerns me that if you're using it as a crosscutting measure that you in a rehab setting there's an intense level of rigor. Many rehab facilities have purchased additional software with branching logic that helps them score the FIM.

All of that will not, I mean, they would have to pay for these in other settings that aren't currently using the instrument. And you will get a very different, I would think, level of assessment from an OT assessing self care as opposed to a nurse or a nursing aid assessing self care.

So those are just some concerns that I have with the level of rigor and the level of

assessment I think you'll see variations in.

DR. CLOHAN: I think from a clinical point of view that's always going to be a challenge unless you are willing to specify exactly what clinician or staff person or technician will perform measures. And so I think a number of measures don't specify who will actually perform them.

And it goes both ways of whether the more rigorous end up with higher scores, lower scores. What we're looking for accurate and consistent scores. You know, we've measured reliability and validity on that to show distinction.

Some level of training is going to have to be required of any measure that's adopted. But you raise good questions. All of these things can go wrong. But to try to, we've tried to come up with a measure that is as reliable and usable as possible and I think it's done that and they're variations of types of facilities where it's used now, and it's

successfully being used by different clinicians.

So I feel very optimistic that that's been well demonstrated.

MEMBER BRADLEY: And I appreciate Lee's comment about the amount of training that may be required and how UDS is held to a standard for making sure that their users go through a training credentialing because we've seen these measures over the years being used for quality and comparing facilities and now for payment.

I think it is important. I think that's a question we should raise on any of the measures that come forward is how much training burden and consistency is there among these tools and among providers using these tools because they are, ultimately, many of them are going to be used for payment and there are ways to game the system if there is not some type of oversight on how clinicians are trained or people administering these tools are trained.

So I guess it's more of a question for CMS in terms of why would there not be that

requirement for some of these measures that could end up in the payment world.

CO-CHAIR STILLE: Lisa?

MEMBER SUTER: Thanks. I guess my question about usability is what information is actually given back to the facility to help them use information to understand their performance?

Obviously they're submitting the data so they have the individual patient level results. But in terms of comparative data, what does the score look like? Are they given their expected rate, you know, their actual to expected ratio? Are they given national results or state results, a spectrum?

You know, I'm still struggling how to interpret the value. And so it would be helpful to know what information the hospitals, the facilities, excuse me, get back. And if you mention that somebody was using this information, how are they using it.

An anecdote would actually be really helpful in terms of wrapping my brain around

that. Thanks.

MEMBER BIERNER: And I'll try to just tell you something practical about its use in the IRF is that there's something that's not mentioned here is something called FIM efficiency where you look at change in FIM over length of stay.

And so we get data as being a participating member with the UDS system from a regional level. You can look at it at a national level. So you can compare your results to various other larger entities and see where you are.

And by looking at things like FIM efficiency, you can see is it an issue of my length of stay or is it more, you can look also at your level of intensity of the type of patients, things like that.

So there's a lot of reporting that is available, at least in the IRF. Whether that will be available in this new setting, I think it remains to be determined.

MS. DANN: I just want to reiterate one thing if I could two, three questions ago. The measures are derived from the FIM instrument which has been around for 25 or greater years.

And we have created things which we have given to CMS over time. When you mentioned branching logic, we do flow charts. They're a part of the training manual. And we are willing to continue giving the refinements to whomever.

CO-CHAIR STILLE: Sherrie?

MEMBER KAPLAN: I just have a concern about the relationship between reliability, validity, and usability because if a measure can be unreliable but it's usually when it's unreliable it's not valid.

It's very unusual to have a measure be unreliable and valid. It can be consistent and wrong like my bathroom scale. I love my bathroom scale, it's consistently wrong. But it's very unusual to have it be inconsistent and right.

So what we just did was kind of said okay, this is inconsistent and it's right. And

then the question is can you use it under those circumstances. And I think, you know, I think we may have not been thinking along those lines.

But it raises the issue of then can you use this measure with any confidence that you're going to do something -- that's why I raise the issue of how responsive is this facility to their efforts to try and manipulate, to try and improve their performance.

Is this responding to quality? That would have helped me a lot, have more confidence that this is actually valid and therefore usable by facilities to act to improve quality.

DR. NIEWCZYK: So to answer the question, the FIM is very reliable. It's been heavily published on the stability, the intra rater, the inter rater, and it's been extensively tested.

So it is reliable. Our Cronbach's alpha was very strong. Our intra item correlations were all high, you know, highly correlated and statistically significant.

So I don't know where the assumption is that it's not reliable. Maybe, you know, it's tracing back to the ICCs that we were asked to provide, we were hesitant. But if we know what you want, we will give it to you.

But yes, it's a reliable measure and it is highly predictive. It's predictive in terms of function, it's predictive in terms of discharge location.

And additionally, what we do provide facilities is what we call the PEM report. So it's a program evaluation model. And it takes into account the average patient's functional change as well as discharge disposition and it also adjusts for the facility case mix, so the type of patients that are within the facility as well as if it's a freestanding or a unit. So there's other variables that go into this model as well so we're able to rank outcomes.

MEMBER KAPLAN: Yes. I just want to be really clear. I'm not talking patient level, it is absolutely reliable. It's the facility

level now that we're worrying about because that's the unit being compared.

So we kind of left behind the patient level reliability which is absolutely totally on board with everything you said. The facility level reliability we're not, we're now struggling with.

DR. NIEWCZYK: And it's extremely rare to find a low performing facility in one quarter jump to the best performing facility in the next quarter. So what we find is that the high performing facilities are usually relatively stable, the mid, you know, the average and then the low stay pretty consistent over time.

There's sometimes a bump up or two but you don't see some of those really drastic changes without having some other change occur like administration or something of that nature.

CO-CHAIR STILLE: Okay, great. Jennifer and then Linda. And then I think we need to -- oh sorry, okay.

MEMBER MELILLO: So the research that

you're citing was done on --

(Simultaneous speaking.)

MEMBER MELILLO: -- correct?

DR. NIEWCZYK: Yes.

MEMBER MELILLO: First level patient?

DR. NIEWCZYK: We have published papers on IRF patients as well as SNF patients. So it's not exclusive to IRF.

MEMBER MELILLO: Okay. But did all of those undergo the same training and --

DR. NIEWCZYK: Yes, they would all have passed a master exam if they're used in our repository. Yes.

MEMBER MELILLO: Okay. So, in that case, it's apples to apples.

DR. NIEWCZYK: Correct.

MEMBER MELILLO: But in order for that to stay apples to apples, then the SNFs and the LTACs would have to invest in that same level of training?

DR. NIEWCZYK: Yes, if they wanted to be included in our repository. If they wanted to

use the tools, the instrument and the training guide and materials, it would be at their own, you know.

MEMBER MELILLO: And that would be a subscriber fee?

DR. NIEWCZYK: Not for use.

MEMBER MELILLO: For yours?

DR. NIEWCZYK: But yes, for the services, correct.

MEMBER MELILLO: Okay, thank you.

MEMBER BRIGHT: So I have two questions and probably the first one piggybacks off of what Linda was just asking. So walk me through a facility that's not a subscriber. I mean okay, they have access to the tool which they have to use as is for free. And they can take advantage of your published training guide, but on their own.

Where does that information go? Then are they on their own to determine what that says about their facility? Is that information that they're submitting to CMS? I guess what I'm

failing to see is the connect point between a facility that is currently not using any kind of valuation and suddenly they have this quality measure tool available to them. But then what?

And is it -- I understand the free thing, but by definition it's kind of forcing everybody to be a subscriber in order for it to be of value. That's what I'm hearing, so disabuse me of that conclusion. That's why I'm asking the question.

MS. DANN: That's not the case. Go ahead.

MEMBER BRIGHT: Okay. And then my second question is kind of following on that which is, is it important -- I'm hearing a lot of concern about inter-facility comparison and being able to use this tool and come out the other end with a picture of what it looks like in one skilled nursing facility but then being able to compare that with 20 others.

So does there need to be accompanying guidance to this quality measure that talks about

its appropriate use because I'm concerned like other members are that we're using the word use and it could come out a lot of different ways, and there's nothing here that says it's appropriate to use it this way, it's not appropriate to use it this way.

So, sorry, two questions.

CO-CHAIR PARTRIDGE: I think I'm going to forestall that a little bit because Jennifer's a new member and the issue that you raised gets raised around this table all the time. To what extent should the NQF endorsement be accompanied by recommendations for how the measure is used. Is it a QI measure, is it an internal reporting measure and so on.

Theoretically, everything that we endorse is supposed to be able to be used for public reporting. I think that's become to some degree somewhat, I don't want to say soft. But as we have studied the measures, we aren't always saying, is it ready to be publically reported right now.

So help me you all. I don't want to be the only person saying this. But it does come up time after time throughout NQF standing committees. We need to keep talking about it.

MS. DANN: And I want to clarify what, if the measures are endorsed and adopted, our offer to CMS is to be able to use the measures royalty-free in any venue they choose. So if they choose to report on them in their quality of reporting, we will have given them everything they need to do that

DR. NIEWCZYK: Or an industry as a whole. I mean, certainly they'd be at, you know, be within reach.

DR. CLOHAN: And all sorts of quality measures can be useful within your site or within your organization, even if you don't have national benchmarks.

I mean, even something as simple as hand washing, for example, people all over like to use sort of a consistent methodology of how you score hand washing so that then within your

own site you can make improvements based on that, even if you don't have access to other people's hand washing data.

And so functional measurement can serve that same purpose. How can you identify which functional areas were strong in achieving improvement or were avoiding any diminution of function in those patients who aren't capable of improving.

So there's value to it, I think, even if you don't have the broadest array of comparison information.

CO-CHAIR STILLE: Okay. Are we ready to vote? Oh, I'm sorry. I didn't see, one more --- oh, Lisa, down there. No, everybody? Okay, false alarm. Okay. Usability and use.

MS. QUINNONEZ: Voting is now open for usability and use of Measure 2769, which includes accountability, transparency, improvement, progress demonstrated, benefits outweighed, evidence of unintended negative consequences. And the option number 1 is high, option number 2

is moderate, option number 3 is low, and Option Number 4 is insufficient information.

Okay, it looks like all votes are in. Voting is now closed. The results are 16 percent voted high, 58 percent voted moderate, 11 percent voted low, and 16 percent voted insufficient information.

CO-CHAIR STILLE: Okay. So overall suitability of 2769. And then we'll have a discussion about the ones that are like it in a few minutes.

MS. QUINNONEZ: Voting is now open for overall suitability for endorsement of measure 2769. Voting option number 1 is yes, voting option number 2 is no. Okay, all votes are in. Voting is now closed. The results are 84 percent voted yes, 16 percent voted no.

CO-CHAIR STILLE: Okay. So, good. So similar to our discussion previously today, 2774 and 2775 may be enough alike that we can kind of put them together. But I think we need to have a little bit of discussion, just to make sure there

are no major issues if we do decide to do that. Anyone, particularly the primary reviewers for the other two, Peter?

MEMBER THOMAS: With respect to the motor measure and the mobility measure, the motor measure is inclusive of all the same characteristics as the mobility measure. And it seemed quite redundant. Why would you have two separate ones?

The obvious answer is because maybe all you want to measure is mobility. And you don't want to measure the other things. But it just seemed redundant. Can you explain why we should approve two when we can approve one?

DR. NIEWCZYK: Yes. So the motor is really encompassing both self care and mobility. So that's the sort of gross, you know, functional assessment. And then the other two are, you know, sub-domains.

So certainly if a patient has, you know, very restricted mobility, and that's not going to be functional goal, then it might

behoove the facility to use the self care measure.

Perhaps they may want to use both, but the self care would probably be a better assessment in terms of change from admission to discharge. It's just providing various levels of functional measurement for the facility's use.

MEMBER THOMAS: It's currently not required, right, it's all voluntary right now. If they were ever to require, CMS, for instance, were ever to require something along these lines, my presumption is that they wouldn't have you do, you know, duplicate efforts in that way. So this is almost just kind of protecting the different domains. And maybe you want it all in one, or you want it broken out.

DR. NIEWCZYK: Exactly, yes.

MEMBER THOMAS: Okay.

DR. NIEWCZYK: It wouldn't be duplicate data collection, so it would be looked at only as a composite score if it was done in its entirety.

MEMBER THOMAS: Great.

DR. NIEWCZYK: Yes, same elements.

CO-CHAIR STILLE: Beth, you're next.

MEMBER AVERBECK: This would have applied to the previous measure. So it's more just a question of understanding if it changes what we would do with it.

You did do some stratification on socioeconomic and the demographics. I couldn't tell if anything looked like it was statistically significant or not. It didn't appear that it was in looking at it. And so I appreciate you taking a look and at least reporting that. But just for the discussion, if this applies to all the measures it would at least worth noting.

DR. NIEWCZYK: Yes. I don't believe there is. Maggie, you can speak to it if I'm misspeaking here.

MS. DIVITA: No. I didn't actually do any statistical testing with the large case count. And we always are a little bit wary of looking at p values. But I can certainly do that

if that's something you need.

DR. NIEWCZYK: In terms of what the averages are, they're almost equal, so yes.

MS. DIVITA: Yes. They're pretty close across, yes.

CO-CHAIR STILLE: Other comments? Peter?

MEMBER THOMAS: Seeing no others, I'll take the opportunity to gratuitously volunteer that I like the measures. There's obviously a very long history in data collection on all these elements.

But these really stop somewhat short of the longer term outcomes. And now that we're getting into measuring outcomes, and functional restoration, and functional improvement, and status, I think it's really important that we begin to measure things that people really care about a lot, like the ability to live independently, the ability to return to work, the ability to be active in the community, and to return to their previous leisure, recreational,

athletic activity.

All those things that people really care about in terms of long term, functional restoration or status, I think is really critical. So obviously not completely relevant at all to this but just on a broader issue, broader scale, I would encourage you to continue to do more work in the area to try to get into those areas. Because they're really important.

DR. NIEWCZYK: Agreed.

CO-CHAIR STILLE: Thanks, great comment. So I think there seems to be consensus, please let me know if there isn't, about considering 2774 and 2775 just in terms of their overall suitability, having had the discussion about 2769 previously. Okay, let's do it.

MS. QUINNONEZ: Okay. Voting is now open for the recommendation for overall suitability for endorsement for Measure 2774. Option Number one is yes, Option Number 2 is no.

Looking for two more, three more votes. Did everyone on the phone submit their

votes?

Can we have you submit one more time, just click one more time. We're missing two votes. Okay, then we're good. Okay. Voting is now closed. All votes are in. All votes are now in, and voting is closed. The results are 79 percent voted yes, and 21 percent voted no.

Voting is now open for the overall suitability for recommendation for endorsement of Measure 2775. Option Number 1 is yes, Option Number 2 is no. All votes are in, and voting is now closed. The result is 100 percent yes, zero percent no.

We are going to re-vote that one, one second. Okay. I'll have you resubmit your votes for the overall suitability for endorsement of Measure 277, Option 1, yes, Option 2, no. All votes are in, and voting is now closed. Okay. The results are 79 percent yes, and 21 percent voted no.

CO-CHAIR PARTRIDGE: Home stretch. Next up we have, again, our colleagues from

UDSMR. We're moving to the other set, which is related to long term acute care facilities, one, two, three measures beginning with 2776, and then 2777, and 2778. And so the floor is now open for discussion. Let's start again with 2777, which is the self care.

MEMBER BIERNER: Well, I was one of the leads on that one.

CO-CHAIR PARTRIDGE: Good.

MEMBER BIERNER: I think it's very similar to the issues that have been raised in the previous, just to point out that the LTAC, this is really a new setting for its use. And so I don't really, I'll ask the UDSMR to tell us, but I don't think there's as much data, if any, from LTACs. They said 23 LTACs, I think is what you had. Is that right?

DR. NIEWCZYK: So I agree in terms of our repository for the long term care facilities. It is our smallest data site, and it is still -- use of the measures are still in its infancy in this post-acute care setting.

However, the files that I had submitted for the measures did include almost 4,000 patients, so 3,922 patient level data points from six facilities. In the ICC, we exploded that out to 16 facilities. So we do have, you know, a respectable number of subscribing facilities, but it's nothing compared to our 800-plus IRFs or, you know, almost 200 SNFs. So yes, I agree.

MEMBER BIERNER: And I found, I put in my comments, in terms of the patients that are ventilator dependent, you quoted a study on your website that about 28 percent of ventilator dependents were discharged to home. And is that --- I'm just concerned about -- what do you think the sensitivity is to detect changes, given the level of some patients like that that might be a much lower level?

DR. NIEWCZYK: So often we don't see as drastic functional improvement as we do in other post-acute care venues. However with, you know, measurement of function in mind, there is

some slight improvement.

And certainly, you know, we can assess maintenance or those that are, you know, starting to decline and likely to be readmitted to acute care or other intensive care. So there is some utility still to assessing patient function.

And again, if we take this in mind with patients measured across the post-acute care continuum, it is noteworthy to know what they look like in the other venues prior to admitted to the, you know, subsequent facilities.

The rating scale is seven levels. So it's from one to seven. So there is some, you know, good degree of sensitivity within the functional assessment ranges.

MEMBER SUTER: So I agree with Samuel's comments. I was also a lead discussant on the self care measure. And I have very much the same concerns as with the earlier measures, particularly around usability, particularly for this patient group and particularly the fact that it's a change measure.

I'm wondering how the information is presented to the facilities, and how that's useful, and whether or not it's broken out so that they can understand how to use the information, I guess.

DR. NIEWCZYK: So the same, you know, the same protocol for long-term care as it would be for skilled nursing or for IRF. So if they are subscribing facilities, they have access to all of our materials and credentialing. Certainly they'd have use of those materials, even if they weren't subscribers.

But I believe that there is utility for measuring function for patients even that are medically complex and need, you know, stability. And we do capture at the lowest level which is complete dependence. So we can even see small, incremental changes that can be captured.

CO-CHAIR PARTRIDGE: Stephen?

MEMBER HOY: Are we doing 2777 or 2776 still?

CO-CHAIR PARTRIDGE: We're beginning

with 2776.

MEMBER HOY: Okay.

CO-CHAIR PARTRIDGE: But that's fine if you --

MEMBER HOY: I'm going to hold then.

CO-CHAIR PARTRIDGE: Well, no, don't. Go ahead. Because quite possibly it's general.

MEMBER HOY: Well, I was a lead discussant on 2777.

CO-CHAIR PARTRIDGE: That's okay.

MEMBER HOY: And I did notice possibly a shrinking gap in care. Their numbers in this set stops at 2011 data set, I know. But I know you had some limited data. And I don't know what we're supposed to assume about trends, but I just wanted to call attention to the shrinking gap of facilities performing above expectations going into 2010, 2011.

And I know we can't really speak too much to the distribution yet of the samples, but could you speak a little to that trend or if you've seen anything else that might suggest a

shrinking gap?

DR. NIEWCZYK: Yes. I mean, I would love to think that there's a shrinking gap. I think this is just an artifact of small samples. So when we have it at the facility level, you know, the patient count samples are respectable, nearly 5,000 and sometimes well over that.

However, when you're looking at the number of facilities, you know, ten versus three, so you're going to see some, you know, slight differences. But the percentages really, you know, exaggerate those differences.

MEMBER HOY: Which is interesting though in the LTAC utilizations kind of trending up right now.

DR. CLOHAN: I'll just a make a clinical comment. It harkens back to the committee's discussion much earlier today on measurement of pain and the question of the importance of asking the question if you want the care to change.

And so I think LTACs have not

traditionally measured function. And so I think there's logic behind believing that asking questions about function could, in fact, lead to significant improvements in care.

I would use, as an example, if you're not thinking about the importance of helping a patient become more independent in transfer, say moving from bed to chair, whatever, and you don't think about it, it becomes very easy just to do that function for the patient, not try to help them become more independent.

And we know there's good data to show the more a patient can move around, the less they're at risk for pressure ulcers, pneumonia, other things. So I think that the mere asking of the question could be a good leverage to improve on the quality.

CO-CHAIR PARTRIDGE: Actually, this is a question of mine for Sam. What's the case load like in the long term acute care facility, typically? Who are your patients?

MEMBER BIERNER: So they're usually

medically debilitated persons that have had a long hospital stay and require, you know, intravenous therapy or drug therapy. They may require respiratory therapy, ventilator dependency, things like that. So they're a fairly debilitated group of, well, my term, medically complex patients.

CO-CHAIR PARTRIDGE: That's what I kind of thought, but I wasn't sure. And this is a new, more or less, new venture for you too. And I gather, therefore, we've got a really significant small numbers problem. Yes.

Are we ready to vote on the evidence question? I think I am. Peter?

MEMBER THOMAS: Well, I did ask that question earlier in the day about how you could take a measure set that was primarily developed around IRFs and apply it to SNFs, and then LTACs, and was that doable.

And you tended to say yes, because what you're measuring is similar across those settings. And so it may be a small measure set,

but you didn't seem uncomfortable with it then. Is that --

DR. NIEWCZYK: So if you have a facility, and you have providers that did improve function, but you didn't measure it, and then they were discharged and maybe admitted downstream to an IRF or to a SNF, how are you going to capitalize on the changes that you've made for that patient if you don't measure it?

Clearly, they just might maintain. And maybe for the majority that's what happens. But we did see some change within patient function from admission to discharge. So there are LTACs that are completing the assessments on patients that are making some functional gains.

DR. CLOHAN: And the measurements, the technique of measuring function is consistent across venue. You're looking at, you know, can the patient take steps, how much assistance do they need to move from a wheelchair to a toilet, those kinds of things.

And so the measurement technique is

consistent. And these things being measured, just like a person, who once they're discharged out of post-acute care they're interested in the things that you talked about, return to job, access to community, recreation, those kind of things.

When you're at LTAC level, and inpatient level, SNF, whatever, you're very interested can you walk, can you get to the bathroom, can you feed yourself, those things.

DR. NIEWCZYK: Eat, yes. I mean if -- come in with a feeding tube and then were discharged eating independently, even with some modifications, I mean, that's a big improvement.

MEMBER THOMAS: Last comment. Yes, my understanding of LTACs is that there is some variation in them. And the ones that come to mind are the ones that are kind of the premier LTACs to me, Craig Rehabilitation Hospital in Denver, Colorado, or Shepherd Center in Atlanta, Spaulding in Massachusetts. You know, they provide significant rehabilitation services to

people with brain injury, spinal cord injury.

That's not an endorsement of those individual places, they're examples. But they certainly provide rehabilitation to improve and to, you know, to address functional status, for sure, so for what it's worth.

CO-CHAIR PARTRIDGE: I think, Peter, my question wouldn't be so -- would be if you've got -- all you know is something about six facilities.

It's kind of hard to sit back and say, well, we're comfortable with a significant variation in performance among six facilities. That's all, as opposed to the other two categories we've discussed here. On the other hand, that doesn't mean it isn't valuable to start.

MEMBER BRADLEY: Then my comments may relate more to usability, but since we're talking about change, there is another element to this scale, as well as in terms of how we use it in a hospital.

Even if there is no change in a patient, it measures, it kind of goes back to more of the earlier research that was done. It measures the burden of care a patient requires.

So even if a patient doesn't make any functional change, it provides providers the information that they need to train the family, to teach the families, and to do the education needed if these patients are projected to go home.

So it provides information even if it doesn't measure any change as useful information. And I think sometimes we forget that we're using it for other purposes other than just measuring did the patient go from A to B.

MEMBER FRIEDMAN: Yes, I just have a question about -- in any of these measures and questions, is there anything that addresses caregiver stresses? As a lot of these patients that are going home will likely have a caregiver who need to support them a great deal.

DR. NIEWCZYK: It's a great question.

It does not directly measure caregiver stress, however there is published studies whereby the functional status that the patient is at, in terms of where their, you know, total score is, equates to a number of hours of one-on-one helper care.

So it can be useful for communications with families and caregivers in terms of, you know, instead of mom's going to need a lot of help, it quantifies what that lot of help would be, meaning eight or more hours per day versus one to two hours per day.

CO-CHAIR PARTRIDGE: If there are no more comments, and in view of the time -- I'm sorry, Linda, I don't mean to cut you off. Go ahead.

MEMBER MELILLO: So just very quickly, my concern is that, while there were six LTACs, three of them were in Massachusetts. So I think that's very narrow. Also, did you collect your data pre or post 25 percent rule?

DR. NIEWCZYK: So this, when was our

first step -- so there would have been, you know, pre-75 percent rule. Absolutely.

MEMBER MELILLO: And then just a general comment. It concerns me that quality would be assessed in LTAC populations that may not include measures that are already existing for IRF. So like swallowing, vent weaning, you know, that those would also be very good measures to include --

DR. NIEWCZYK: Absolutely.

MEMBER MELILLO: -- in a quality of LTAC service line. And so I just have that concern that, you know, by limiting it to existing measures, that we may not be capturing the essence of what is an LTAC.

DR. NIEWCZYK: Absolutely. This is not an all-encompassing measure. This is only looking at a narrow focus of function.

And I just want to add that the data we supplied for you is from six SNFs, but we don't only have six SNFs in our repository or, I'm sorry, LTAC. I'm on previous --- you know,

in the ICCs we had included 16 facilities.

We do have much more data. We initially had compiled this at roughly 5,000 patients per venue when we submitted for Phase 2 measures and had all sites included in our applications. So that's why it's looking like it's only six. Rest assured, there is more data behind the scenes that we could provide if you would like.

CO-CHAIR PARTRIDGE: Are we ready to vote on evidence with 2776?

MS. QUINNONEZ: We are now voting on measure 2776, functional change, change in motor score, and long term acute care facilities. Rationale supports the relationship of health outcome or PROs to at least one healthcare structure, process, intervention, or service.

Option Number 1 is yes, and Option Number 2 is no. Okay, all votes are in. Voting is now closed. The results are 95 percent voted yes, and five percent voted no.

CO-CHAIR PARTRIDGE: Moving on to gap,

et cetera.

MS. QUINNONEZ: Okay. We're voting on the gaps. Performance gaps, data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups.

Option Number 1 is high, Option Number 2 is moderate, Option Number 3 is low, and Option Number 4 is insufficient.

All votes are in, and voting is now closed. The results are 11 percent voted high, 37 percent voted moderate, 21 percent voted low, and 32 percent voted insufficient.

Voting on reliability for measure 2776, reliability including precise specifications and testing. Option Number 1 is high, Option Number 2 is moderate, Option Number 3 is low, and Option Number 4 is insufficient.

Looking for three more votes. All votes are in. And voting is now closed. The results are 11 percent voted high, 42 percent voted moderate, 16 percent voted low, and 32

percent voted insufficient.

Are we ready to vote on validity? We're voting on validity for Measure 2776, validity including specifications consistent with evidence, testing, exclusions, risk adjustment, stratification, meaningful differences, comparability, and multiple specifications, and missing data. Option Number 1 is high, Option Number 2 is moderate, Option Number 3 is low, and Option 4 is insufficient.

All votes are in, and the voting is now closed. The results read five percent voted high, 53 percent voted moderate, 21 percent voted low, and 21 percent voted insufficient.

Feasibility of Measure 2776, data generated during care, electronic sources, and data collection can be implemented. Option Number 1 is high, Option Number 2 is moderate, Option Number 3 is low, and Option 4 is insufficient.

We're waiting for two more votes. And we have them. All votes are in, and voting is

now closed. The results are 21 percent voted high, 58 percent voted moderate, 16 percent voted low, and five percent voted insufficient.

We are now voting on usability and use for Measure 2776, accountability, transparency, improvement, progress demonstrated, benefits outweighed evidence of unintended negative consequences. Voting Option Number 1 is high, Voting Option Number 2 is moderate, Option Number 3 is low, and Option Number 4 is insufficient information.

All votes are in. And the voting is now closed. The results read 11 percent voted high, 47 percent voted moderate, 16 percent voted for low, and 26 percent voted for insufficient information.

We are now voting for the overall suitability for recommendation for endorsement of Measure 27 --- this should actually be 2777, six, 2776. I'll change that. Option Number 1 is yes, Option Number 2 is no.

Missing one vote. All votes are in,

voting is now closed. The results are 58 percent voted yes, and 42 percent voted no.

MS. THEBERGE: Operator, can you open the lines for public comment please?

OPERATOR: Yes ma'am, at this time if you wanted to make a comment, please press star, then the number one. There are no public comments at this time.

CO-CHAIR PARTRIDGE: So 2777 and 2778 remain on our plate. And again, this one's a little awkward.

DR. NIEWCZYK: Well, if I can respectfully ask, you know, we did supply a great deal of evidence related to reliability and validity. You know, certainly had we sat here with, you know, 100 cases or face validity only, I can understand the outcome. But we not only had, you know, concurrent validity, but we also had predictive validity.

To see our square values, you know, in the range that we had, that's highly respectable. Certainly some published literature, you know,

include in our square of 0.17. And it gets published. So I think I'm confused in terms of what the committee wanted for reliability and validity.

MEMBER THOMAS: So the last time, earlier today, there was a discussion about we have, like, two months to provide additional data and to try to work with the staff and figure out whether we can --- some of the, you know, we have a second meeting and talk. Is that the same process that would be used in this instance?

CO-CHAIR PARTRIDGE: Are you suggesting you might make a motion to defer consideration?

MEMBER THOMAS: Oh, I wasn't sure that was the deferral. I thought the deferral was, like, come back to us and we'll talk ---

CO-CHAIR PARTRIDGE: Well, it is.

MEMBER THOMAS: -- talk about this next year.

CO-CHAIR PARTRIDGE: It is. As opposed -- if I'm understanding the discussion

earlier correctly, and I may not, so please, Suzanne and Sarah, stop me. We were really addressing time frame.

If the time frame is pretty compressed, if we say essentially, during the public comment period and so on, if there's additional data you would like to submit, we will consider it right as part of phase, this ---

MEMBER THOMAS: This phase.

CO-CHAIR PARTRIDGE: -- this segment of Phase 3. If you think it would take more time to make the committee comfortable with a recommendation to endorse, say up to a year, then that would be a deferral. Am I right?

MEMBER THOMAS: I'm just suggesting that we heard is this six LTACs, and you said no. We have data on much more than that. And you do ---

DR. NIEWCZYK: I mean, how much is enough? Like, I don't know if it's sample size that you're looking for, you know, certainly I can pull, you know, data from when we first had

an LTAC start collecting.

Sometimes what happens is the larger your database is you have minutiae turn up as being statistically significant when it's not clinically meaningful.

So in an attempt to counter that we, you know, we took a sample and we tried to have the samples equal between settings. Otherwise, clearly our IRFs would have dominated.

So if you took a look at, you know, the submission files, they're all roughly 5,000 patient counts. Of course, when you have 800 and some-odd subscribers in IRFs, there's going to be more variability of those numbers of facilities versus LTACs. We may have a smaller pool, but nonetheless, there were more than six. So if it's just a larger sample you want, that could easily be done.

MS. SAMPSEL: So, Peter, your question is a process question as well. And so basically what happens, there were a number. So gap was consensus not reached, reliability, validity, and

then use, usability and overall.

And so it does go back to the process of, you know -- and I think what Paulette's trying to find out from all of you all is what do you want to see.

You know, they don't have to provide all their data. I mean, we know if health plans test they probably test in three health plans. So kind of, you know, what are those things? And I think that's what we'd like to draw out. Right now, what are those things that, in this eight week period, UDSMR could bring back to you.

Because you will re-vote after public comment on these measures. Did I, is that what you're looking for?

DR. NIEWCZYK: Absolutely. And it's just, I've been involved in this process for quite some time, not only for our measures but, you know, I've been on the line for other measures under consideration.

And it seems to be a great deal of variability between what's submitted for

measures. So sometimes the committee is perfectly fine with, you know, one setting, you know, a handful of -- and other times it's -- so if there's some consistency in terms of what you want, you know, we'll try to provide that.

MEMBER BRIGHT: Really briefly, I think one criteria that really weighed in for me was the fact that this is a new setting for this, okay.

So I think that, to some degree, what you're getting from the committee is a sense of, yes, give us a little more data. Because this is a setting that hasn't done this really. And it's really great that you have such a robust data set.

But you're talking about a data set of organizations that already have resources and a focus on this. And we're talking about trying to pass a measure that could be applied to settings that aren't doing this at all or definitely aren't doing it well.

So I guess, just speaking for myself,

I feel okay with saying could we see some other examples, just prove to us that it's not just six facilities. But that's really it. And the newness of the setting was really my litmus test.

CO-CHAIR PARTRIDGE: Lisa and then Sherrie.

MEMBER SUTER: So for gap and for usability, the information that would be really helpful to me would be looking at facility level distribution of results and some information about the data that goes back to facilities.

I know we heard from Samuel that there's a lot of comparative data, but a little bit more information about that, which it sounds like would be pretty easy to pull together, I think would be very helpful from my ---

DR. NIEWCZYK: Thank you.

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: Yes. Full disclosure, I'm on the National Advisory Committee for Physician Compare. And all of these issues are being raised at the, you know, how do you measure

physician level performance, and how do you discriminate, and at what level should you discriminate.

And there's a fairly robust literature out there at the physician level which is applicable for this unit of comparison. It's a unit being compared. And the issue of intraclass correlation coefficients comes up there. And there's fairly generalized estimation equation. So there's a fairly robust literature out there.

So the kind of facility level comparisons that we were looking for in the reliability, and then validity at the facility level for discriminant validity between facilities, which is what you want if you're going to be the facility, and you're going to either be inappropriately or unfairly compared to folks with different, either sample size distributions or --- you really -- this is not a trivial issue anymore.

And it brings up the issue of usability as well. Because if CMS is going to

then begin to use these for, you know, penalizing people or incenting people, in one way or another, for meeting or not meeting these benchmarks, then the standard goes up.

And so I think that's what NQF right now, Sarah, unless I'm wrong, is dealing with. There is some changing in the uses of these measures now and the potential uses of these measures, even though we're supposed to be agnostic to use, that's starting to move the psychometric properties that we're going to be expecting from these measures.

So I'm still struggling with the issue of what's fair to hold developers accountable for. Because the Betty Crocker of how to do this is in the literature. It's not like we dreamed it all up in this committee. There's a fair amount out there that would be useful and helpful in how to actually do this.

The question is whether it's fair right now to hold measure developers accountable for meeting the requirement if, in fact, the

guidance hasn't been clear enough. That's what I'm struggling with.

CO-CHAIR PARTRIDGE: I think you have heard several of these measures that it would be, it would perhaps raise the comfort level of a number of members if we had the distribution data that Lisa Suter asked for. And you were nodding, which suggests to me that perhaps that's --

DR. NIEWCZYK: That's definitely possible. Yes, absolutely.

CO-CHAIR PARTRIDGE: -- possible. So that would be a suggestion that I would also support.

MEMBER KAPLAN: And the intraclass correlation coefficients at the facility level in the way we were talking about, not the way it was done.

MEMBER THOMAS: Can NQF staff provide some technical assistance or communication with them in order to run that data?

MS. SAMPSEL: No. We won't run the data for them.

MEMBER THOMAS: No, no, no. Not to run it, but TA, technical assistance to do it the right way.

CO-CHAIR STILLE: Yes. We're going to work with Sherrie on that.

MEMBER THOMAS: Yes, okay.

MEMBER MELILLO: Yes. I would like to see the training manual to see what the burden would be to those staff who are coming at these measures new, how complicated is it.

You know, what's the realistic expectation that, within an organization, they're going to be able to use this material to actually score the FIM in a consistent manner as the IRFs are currently scoring it. In addition, I would love to see a sample size of at least 30 facilities.

DR. NIEWCZYK: So is that, like, a rule? I mean, because I just see that that there are some different rules for different submissions.

It seems, you know, like, nowhere did

it ask to show any training materials or anything like that. Had it asked for that, we would've absolutely included it. I just want to be held to the same standards that everybody else is being held to.

MS. SAMPSEL: Well, but I could say, within Phase 1 with the CAHPS measures, the question about how CAHPS was administered and all of these similar questions on CAHPS do come up.

So, I mean, we don't -- do we have CMS bring their 500 page volume? No. But some of these questions do come to feasibility, usability, and use. And I think you're hearing more of this, at least that's what I'm picking up, because these are new to the LTACs. So I do think this is a fair question.

DR. NIEWCZYK: And, you know, we'd be happy to provide it. I'm just, I want as much clarity as possible so we can get you what you want.

MEMBER MELILLO: My level of concern is that you mentioned this is a cross cutting

measure for use with the IMPACT Act which then becomes a compliance related measure and also, potentially, a payment related measure. And so because of that, I think it's critical, it's right.

CO-CHAIR PARTRIDGE: I just tested this script with the lady to my left. The question before us is on 2777 and 2778, the two remaining measures for the UDSMR team and us.

What is the comfort level with this committee about going through all evidence gap, reliability, validity, feasibility, usability, and overall, for each of those two remaining measures as opposed to doing what we have done earlier, and saying we believe the results for each of those two would be the same as with this 2776?

And we need only take the overall recommendation for endorsement vote, just on that issue, comfort level, I'm sorry, not for developers. Lisa? Anybody have heart failure if we did it that way? Sure.

DR. CLOHAN: Is there a need to go ahead and record votes for the two? Or if the committee is moving toward just collecting some more information, might you want to just leave those other two unaddressed for the moment and get that information back.

CO-CHAIR PARTRIDGE: I believe as a process question, in order to allow it to go forward for public comment, we need to take the recommendation vote. Am I right?

MS. SAMPSEL: Yes. It's also consistent with what we did for the rest of day this morning. So I think we should still do that, realizing that, you know, it still may be consensus not reached. But I think as a process, correct, we do need to do that.

MEMBER THOMAS: Similar enough measures that we can assume that --

MS. SAMPSEL: Yes. And in fact, we did this in Phase 2 as well. We carried, even in the report, we carried the discussions forward in the report of saying we had similar discussions.

These were the highlights of the discussions.

CO-CHAIR PARTRIDGE: Okay. So Desi?

MS. QUINNONEZ: So we are voting for the overall suitability for recommendation for endorsement for Measure 2777. Option 1 is yes, and Option 2 is no. Voting is now open.

Looking for three more votes. All votes are in, and voting is now closed. The results are 47 percent votes yes and 53 percent for no.

We're voting now for the overall suitability for recommendation for endorsement of Measure Number 2778. Option Number 1 is yes, Option Number 2 is no. Looking for two more votes, one.

MS. THEBERGE: Deb, we need your vote. Thank you.

MS. QUINNONEZ: Okay, all votes are in, and voting is now closed. The results are 37 percent voted yes, and 63 percent voted no.

MS. SAMPSEL: So this one fails.

There's a public comment that we

missed.

MS. THEBERGE: Operator, can you open the line for a public comment please?

OPERATOR: Yes, ma'am. At this time to make a comment, please press star then the number one. We do have a comment from Deanna Hayes.

DR. HAYES: Hi, this is Deanna with Focus on Therapeutic Outcomes. I want to thank you for the opportunity to observe this. These were some great conversations and things that were really thought provoking to me.

I was unclear with 2769 and why it was passed when it did not appear to meet the standards for being a performance measure. I think you probably just addressed this in your most recent comments. But my request for comment was submitted prior to your last discussion. But I was still unclear about 2769 itself, because it was passed.

MS. SAMPSEL: Actually, Deanna, 2769 was -- hold on a minute. You know, so they were

different measures. I mean, from my understanding from the committee, and this is just my interpretation, is those first three measures were the SNF measures where they have been in longer use.

And there was more data provided where, with the LTAC measures which were the last three that were not recommended, as of right now consensus was not reached. But that was a difference in voting.

DR. HAYES: What I heard in the discussion was a committee member described concerns, and it didn't sound like they were resolved. It sounds like the data is there, and it's available to demonstrate validity and reliability for performance. It's just a matter of continuing to work together to find the right analysis.

MS. SAMPSEL: Right. And so during these next few weeks, NQF staff will work with UDSMR, and they will provide that additional data that was not submitted. And therefore, the

committee will re-vote.

DR. HAYES: Thank you.

OPERATOR: And at this time there are no public comments.

CO-CHAIR PARTRIDGE: Peter.

MEMBER THOMAS: But on the third measure of the LTACs that just failed, can we assume that that was an accurate vote? I mean, there were a couple instances today when it wasn't accurate. Why would it have changed? We didn't discuss anything differently.

MS. SAMPSEL: We can definitely re-vote that one. I mean, I have the same questions. I mean, if anybody changed their vote between the two, you know, I guess, does somebody want to explain, or do people want to re-vote?

MEMBER THOMAS: I'd like to re-vote.

CO-CHAIR PARTRIDGE: Okay, let's re-vote.

MEMBER THOMAS: Only because we've had this happen a few times today. And I just don't know why it would have changed.

MS. QUINNONEZ: Okay. We are reopening the vote for overall suitability for endorsement of Measure 2778. Option 1 is yes, Option 2 is no. All votes are in, and voting is now closed. Thirty-nine percent voted yes, and 61 percent voted no.

MEMBER THOMAS: What does that mean?

MS. SAMPSEL: Well, it still means that UDSMR has the opportunity to bring data back to us, and we'll re-vote after ---

MEMBER THOMAS: Okay. Fair enough, fair enough.

MS. SAMPSEL: -- public comment. It's the same results.

CO-CHAIR PARTRIDGE: Any further comments or discussions? If not, I'm going to excuse our USMDR people. We've put you through a difficult afternoon. And I'm afraid you're still going to have to work with us.

And dinner is at 6:00. I know we lose some of you, because you have other obligations. But those of us who are going to be going for

dinner, I look forward to that in 40 minutes. But otherwise, we're adjourned for the day.

Tomorrow morning we begin at, all right, 8:30 again. Breakfast at 8. We will try, and again, we had some pretty interesting discussions today. We will try to have some time for us to think back over the day and a half before we leave.

And one note, we have to leave at 1:30, I promise you. Because I have committed to my husband that I will leave by 1:30. Chris can take over but --

CO-CHAIR STILLE: Okay, that's fine.

CO-CHAIR PARTRIDGE: -- this half of the Chair will be home, but do think of some of the kinds of questions that have been raised today, some of the things that Sherrie said about what is it that their current algorithms dictate that maybe we can't change, but we might want to recommend be rethought, particularly with regard to PRO-PMs and anything else that, particularly from our new members, as you've hit this process

for the first time. How can we do it better, make it work better for all of us? Otherwise --

MEMBER THOMAS: I just want to say that in comparison to the last two times that I've done this, the staff briefings on this were unbelievably helpful. And I just really appreciate that.

(Applause.)

MS. THEBERGE: So the restaurant is at 16th and K, 1600 K Street. And some folks may wish to go back to their hotel and so if you want to just meet us there. And it's called P.J. Clarke's.

(Whereupon, the above-entitled matter went off the record at 5:23 p.m.)

F

J

P

T

W

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Person and Family Centered Care

Before: NQF

Date: 06-06-16

Place: Washington, DC

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R Gross

Court Reporter