National Quality Forum

Moderator: Leslie Thompson
May 16, 2018
7:55 a.m. EST

OPERATOR: This is Conference # 4589948

Karen Johnson: OK. Good morning, everybody. Let's go ahead and get started. I know we're running just a few minutes late but we wanted to give just a couple of minutes for a few people that were coming in the last minute.

My name is Karen Johnson. I'm one of the senior directors here at NQF. And we've talked, all of us have talked to each other on the phone several times now, but it's really fun to get to see faces and actually meet, so thank you, guys, for coming.

What we're going to do this morning, I'm going to ask our co-chairs to say a brief hello to everybody and also our staff. Maybe we'll start with our staff introductions.

Poonam Bal: Hi. I'm Poonam Bal. I'm a senior project manager on this project.

May Nacion: Hello. My name is May Nacion. I'm the project manager.

Miranda Kuwahara: Good morning. My name is Miranda Kuwahara. I'm the other project manager for this (work).

Andrew Lyzenga: Hi. I'm Andrew Lyzenga. I'm a senior director here at NQF.

Elisa Munthali: Good morning. I'm Elisa Munthali. I'm the Senior Vice President for Quality Measurement. I'd like to welcome you all for being here and it's good to see you in person.

Karen Johnson: And I'm going to ask the co-chairs to do a very brief welcome as well. And you may be a little surprised, not so much if you read your email but we have a different co-chair than you may have expected.

We have Karen Joynt Maddox on the phone and Karen will tell you why she's not here today. But then Dave, David – we've got Dave and David -- well, he graciously stepped in for Karen today. So, I'll just hand it off to you guys to start and then we'll go to Karen on the phone.

(David Behrens): Thanks. One of the very first things we had to do is sort out between two Davids, how we were going to distinguish. And so, we did agree on the phone last week that there would be Dave and I'm David which is how my wife calls me and then there's an extended version if she's mad at me. But I will stop with…

Male: With your middle name?

(David Behrens): Yes, it does. Well, we'll stop with the David. No, I'm happy to be filling in here. I was pleased as could be when Karen was name as co-chair originally – I worked with Karen, I respect her very highly and I don't know that I can fill in the same way but I'm happy to do that.

I am certainly pleased to be named to this group in the first place and be part of it. I think the work that we do here is very important as we get into the day's agenda we'll certainly see that.

We've seen it already to have some clarity among ourselves, to developers, to people who use quality measures about what's a good measure, what's a reliable measure, what's a valid measure. I think there are a lot of important things that we can do, very necessary.

So, I'll do the best I can from this end working with Dave to keep us on track and I think that's the name of the job from this end of the table, to draw the resources in the room and get the best possible product.

(Dave Cella): Hi, everyone. Dave Cella. I'm happy to greet you and welcome all of us here. I think this is really an amazing group of people. I've been so impressed with the reviews that you've done and that we've adjudicated together.

We were thrown a lot of work at the very beginning and hasn't really slowed down since then, so thank you to everyone for all that you've contributed.

And the quality of the reviews is really quite high, but we've also noticed that the inter-rater agreement is not always quite high and that's really what we are compelled to focus on is to get a better handle on what we agree on and maybe what we don't agree on.

It's not realistic to think we would accomplish that today, but if we could set the course toward having some kind of document that could go to developers and to end-users that makes it very clear what from our perspective at least what perspective constitutes a good measure that would be quite I think very appreciated by NQF and by the measure developers and providers. So, that's our task in broad terms.

Karen Johnson: And Karen on the phone, you want to say hello?

Karen Joynt Maddox: Hi, everyone. You switched out of two Karen problem for a two Dave and David problem. But I am back in St. Louis. I am 37 weeks pregnant and no longer allowed to travel.

So, I'll also be obviously stepping away for a little bit on the maternity leave this summer. But it has been an amazing experience to be part of this group and I've learned so much already and very much look forward to remaining engaged.

I would echo everything that has been previously said, that expertise around the table that I'm sure I can see in my ahead of you all sitting on the table is pretty spectacular. And so, I'm thrilled to be able to remain part of the group

and look forward to seeing what we all come up with. So, thank you, all, so much for accommodating.

Karen Johnson: Thank you, Karen. We're going to hand it off now to Elisa who's going to walk us through our disclosures of interest.

Elisa Munthali: Good morning again. So, what we're going to do is combine disclosures of interest with introductions and when you were named to this Committee, you received a form where we asked you a number of questions related to any consulting research or grants that is relevant to this work. And so, what we're doing today is asking you to orally disclose that.

Just a couple of reminders before we go around the room and on the phone, we are interested in disclosures of interest of relevant work, not just paid work but also unpaid. We also wanted to remind you that you sit on this Committee as an individual. You do not represent the interests of your organization or anyone who may have nominated you.

And lastly and this is probably the most important, just because you disclose does not mean you have a conflict of interest. We do this in the interest of transparency and openness. And so, what we'll do is start with your co-chairs. I think we'll start with Dave.

Dave Cella. Hi. I'm Dave Cella. I work at Northwestern University. So, the University is my employer. At Northwestern we have a contract with NQF in the incubator program to develop a performance measure for multiple sclerosis. That's an ongoing project that I should disclose.

I am also the sort of chief architect if you will and steering committee chair, western committee chair of the PROMIS measures that are from time to time considered as measures in PRO-based performance measures. Neither of those is a financial interest for me personally but those are the activities that I'm involved in.

(David Behrens): I was going to say Dave Behrens, I have to learn myself. I'm at Henry Ford Health System in Detroit. Up until two weeks ago, I was Director of the

Center for Health Policy, Health Services Research and strangely enough I chair the Neurosurgery Department though I'm not a surgeon.

I stepped down from both those titles. I'm, therefore, kind of quasi retired but I'm still working on projects including things in the area of health policy and quality of care measurements; I'm very much active in this currently.

I don't have any financial interest to disclose. I have made in the past a number of public statements either in writing or verbal presentations in the general area of quality measurement.

Actually, in this room four or five years ago I chaired a committee on the issue of risk adjustment of quality measures on the basis of socio-economic factors so it's sort of familiar to be sitting on this chair again looking at a group of folks. And so, I've done that and have made some statements about that.

Up until last month I was a Commissioner at MedPAC, Medicare Payment Advisory Commission. In that context, I had a number of comments that are on the public record about what's a good or bad quality measure and I guess I don't think any of those represent conflicts but those are just statements I've made and positions I've taken in the past.

Elisa Munthali: Thank you. Karen?

Karen Joynt Maddox: Hi. So, I'm a cardiologist on faculty at Washington University School of Medicine in St. Louis. So, that's my employer.

My disclosure is that I also have a standing contract for a few hours a week with Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation in which I continue to work on some projects related to socioeconomic status and risk adjustment. I don't work for CMS, but I share a box with CMS technically in terms of those contracting hours for my work.

Elisa Munthali: Thanks, Karen. And we'll continue in the room. We'll start with Ron.

Ronald Walters: First, I'm one of those non-PhDs in the group. I have nothing to disclose, no conflicts of interest, which probably says in itself something except I do a whole bunch of things for the NQF and the quality partners whenever they need me, they know they can call me.

And I agree we've gone from forming and we're somewhere between storming and norming, so that's good, that's progress we're moving to the right.

Lacy Fabian: Good morning. My name is Lacy Fabian. I work with MITRE. Through that, I provide consulting with CMS around their quality measure best practices. Aside from that, no conflicts to disclose. Thank you.

Matt Austin: Good morning. I'm Matt Austin. I'm employed by Johns Hopkins University. I do have a contract with the Leapfrog Group to provide them with guidance around measurement activities. I also have some funding through the Johns Hopkins health system and I'm also on a Betty and Gordon Moore Foundation grant around diagnostic accuracy and we are developing a performance measure which we hope to submit to NQF in the fall cycle.

(Eugene Nuccio): Good morning. I'm Gene Nuccio, University of Colorado, Anschutz Medical Campus. We have contracts with CMS for the home health Oasis instrument measurement development system. So, we work with their quality reporting program.

We also have two contracts with the home health value-based purchasing program. And both the implementation and monitoring contract as well as the technical assistance contract. I also have a contract with MedPAC for measure development.

My background is in quality measurement, risk adjustment primarily in the area of home health. I served with David – that was a fantastic experience by the way – and I'm also a member of the NQF MAP post-acute care, long term care committee.

Bijan Borah: Hi. Good morning, everyone. I'm Bijan Borah and Mayo Clinic is my employer. And I honestly don't have anything to declare or no conflict of interest. Thank you.

John Bott: Hi. My name is John Bott. I'm presently with Consumer Reports although the department I work in is going to cease next month, so I very soon will not be working with Consumer Reports in the next couple of weeks.

I have a small time-limited contract with the Leapfrog Group that will just extend until the end of June. I'm on a few CMS committees that will continue even though my status at Consumer Reports will change.

That's the hospital compare star rating TEP, the CMS qualified health plan TEP and the CMS hospital harm performance measure technical advisory group.

Marybeth Farquhar: Good morning. My name is Marybeth Farquhar. I'm the Vice President for Quality Research and Measurement at URAC. I currently advise on the QRS measures set for Booz Allen Hamilton.

I'm also on the QRS measure set development panel for IMPAQ and I'm also on a number of committees for the Pharmacy Quality Alliance specifically for specialty pharmacy.

(Christie Teigland): Good Morning. I'm Christie Teigland. I'm a Vice President of Advanced Analytics, at Avalere Health which is a consulting firm in Washington D.C.

We do have a few contracts with some organizations that are developing measures. I'm only peripherally involved in those as the data person, one of the data people testing the measures – one is with Abbott under a CMS contract looking at some nutrition measures.

We've worked on NCQA measures in the past, specifically the 30-day all-cause readmissions measure which we tested in the Medicare Advantage population; we have a lot of Medicare Advantage data. And a few other overuse measures.

I'm on a Yale expert panel looking at a new measure for diabetes, both overtreatment and under-treatment. And I am also on the Pharmacy Quality Alliance quality measure expert panel and on their risk adjustment expert

panel, now looking at adjusting some of their medication use measures for socioeconomic risk factors. And I am on the National Quality Forum Standing Committee on disparities. That's it.

Susan White: Yes. OK. Hi. I'm Susan White. I am employed by the Ohio State University. I specifically work in the James Cancer Center. I'm the Administrator of Analytics there.

I've been involved on paid work, involved with the Alliance of Dedicated Cancer Centers, helped shepherd through the 30-day readmission for cancer measures sponsored by Seattle Cancer care Alliance and probably will continue to help with those measures as we go forward. I will just opt out of any like everyone else anything we work on, but that's it. Thank you.

Zhenqiu Lin: Hi. My name is Zhenqiu Lin. I work at the Yale Center for Outcome Research and Evaluation. I'm the Director of Analytics. So, we have contract with CMS to develop quality measure for hospital and outpatient setting and so there will be many measures before this Committee otherwise I don't have any conflict of interest.

Michael Stoto: Good morning, everyone. I'm Mike Stoto. I'm on the faculty at Georgetown. I'm also part time at Harvard Chan School of Public Health. I don't have any measurement-related research or administrative responsibility to those places although I do teach something about measurement at Georgetown.

I'm on an NQF panel on prevention and population health and previously some related panels in that area. And I was previously on two different CMS TEPs in related areas as well but those are completed out.

Steve Horner: Good morning. I'm Steve Horner, Vice President of Clinical Analytics for HCA. I don't have any conflicts of interest to disclose but I do serve on the hospital advisory panel around quality measurement for Blue Cross Blue Shield of Louisiana as well as Anthem. And I also serve on the hospital advisory committee for the Pioneers in Quality Joint Commission.

Laurent Glance: Good morning. My name is Laurent Glance. I am a cardiac anesthesiologist and I work over at the University of Rochester. I'm the Vice Chair for

Research there. I also have some secondary appointments in the School of Public Health Sciences as well as an adjunct appointment at RAND.

I have some disclosures. I am involved with some of the quality measurement activities at the American Society of Anesthesiologists as well as at the Society for Cardiovascular Anesthesiologists and the Anesthesia Quality Institute. And I also serve on the NQF Standing Committee on readmission. Thank you.

(Jennifer Perloff): Hi. I'm Jennifer Perloff. I'm a health services researcher at Brandeis University. Our work at Brandeis is focused on developing an episode grouper for Medicare, so we develop resource use measures.

I'm currently involved in a project with the American College of Surgeons where we're laying quality into that episode framework.

And in addition to my measure work, most of my other research is on attribution, so thinking about the relationship between measures providers and entities. And in terms of my NQF secret identity, I'm involved with the attribution work that's happening here.

Elisa Munthali: Thanks to everyone in the room. So, we'll go to the phone. I'll start off with Paul Gerrard. And I have a note here that you'd be off and on, so this might be the time that you're off. Joseph Kunisch?

(Joseph Kunisch): Yes. Hi. Joe Kunisch. Unfortunately, my flight got cancelled so I can't join you in person there. But I'm the Enterprise Director of Clinical Quality Informatics at Memorial Hermann Health System in Houston, Texas.

And disclosures, in that role, my department does a lot of electronic clinical quality measure development and feasibility testing with mathematical policy research. I also sit on the Yale hospital harm technical expert panel and also the CMS measure development TEP. And I also am a committee member on the (HENS) North American Patient Safety and Quality Committee. Thank you.

Elisa Munthali: Thank you very much. Thank you. Jack Needleman?

Jack Needleman: Good morning. Sorry I cannot join you today. I'm Professor and Chair of the Department of Health Policy and Management at the UCLA Fielding School of Public Health.

I have no financial disclosures. I was one of the co-developers of the AHRQ Patient Safety Indicator for mortality among patients with -- post-surgical patients with selected complications.

And I have provided unfunded support to the American Nursing Association in its effort to get its staffing measures re-endorsed and put on the Hospital Compare system. And I'm a member of the NQF cost and efficiency portfolio Standing Committee and its predecessor committees.

Elisa Munthali: Thank you very much. Sam Simon?

Sam Simon: Good morning, everyone. Sorry that I can't be there. Unfortunately, weather cancelled my flight. So, I'm Associate Director at Mathematica. Mathematica holds several measure development contracts with CMS and I am a principal investigator on a contract to develop and maintain eCQMs that are in the (MIPS) program.

Elisa Munthali: Thank you very much to everyone. And I just wanted to remind you if at any time during the meeting you realize that you have a conflict, we want you to speak up.

You can approach any one of us on the NQF staff or any of your co-chairs. Likewise, if you feel that during the meeting one of your colleagues may have a conflict or is acting in a bias manner, we want you to speak up and approach us.

So, having heard all of the disclosures of interest, do you have any questions of me or of your colleagues? It doesn't look like you do. Thank you.

Karen Johnson: And thank you. A couple of people were kind of at the last minute not able to come. So, I just want to acknowledge their contributions thus far -- Sherrie Kaplan and Jeff Geppert.

So, I don't think they're going to be able to call in even today so, I want to send warm thoughts their way for whatever is going on with them. The only other thing I had is did I miss Paul Kurlansky? Is he…

Female: He's still on his way.

Karen Johnson: He's coming. OK. So, we'll have to catch Paul and Paul, the two Pauls on the (DOIs) a little bit later.

Elisa Munthali: Well, to start us off with the meeting objectives so today our goal is really to review the current processes and discuss improvements or changes to that process.

We'll also spend a good deal of our time discussing reliability and validity and hoping to obtain consensus on that conceptual definitions for those. We'll also discuss possible changes to the measure evaluation criteria and we'll finally just review next steps.

And before I turn it over to Karen, just a few housekeeping announcements, if you haven't already logged into our Wi-Fi, the Wi-Fi log in information is on the desk over there to the left. The log in is guest and the password is NQFguest. The bathroom is outside, right outside the double doors to the right.

And if you'd like to speak, just like this so you would just hit the speaker button, place your (temp) card up so that our chairs know to all on you. Please state your name when you're speaking to the microphone for the public participants on the webinar. And then when you're finish speaking, just turn off your mike. And then I'll turn it over to Karen.

Karen Johnson: Thank you. And another thing about these mikes sometimes you have to get pretty close to them and it may not matter so much for us in the room but for people on the phone it really helps to be close to the microphones. So, as I said that I pushed the microphone back so sorry about that.

To get us started today, I wanted to just give everybody a little background. For those of you on your panel, this is probably a review but we may have

some people on the phone who may be aren't as cognizant of kind of why you exist and what your charge is. So, I wanted to make sure that we cover that today.

So, first of all, this idea of having an external group who could help us think about methods came out of our last Kaizen effort that we had here in NQF so we do this on occasion. We do try to improve our processes as we go along. And almost exactly one year ago today as a matter of fact it was May 18th we had this idea of this methods group to come together.

So, really only a year and we've made incredible progress so far. We had that event in May. I June, we started working out internally how we were going to make this happen. Put out a call for nominations in July and had you guys seated by September. So, it was a fast thing and we are so happy and grateful that you applied and that you want to do this work with us.

Our charge – and apologies I didn't ask for the clicker – the charge of the Methods Panel is twofold. First and I think everybody knows this well by now, to conduct evaluation of complex measures for scientific acceptability – oh, thank you. So, that is a job in and of itself and we do thank you very much for that.

We also are asking you guys to serve in an advisory capacity to NQF. And that's what today is about, that's what our monthly calls are about. So, we know that we have a lot of issues that need to be resolved and I think together we'll be able to do that.

Just going back real quickly, evaluation of complex measures, we limited the measures that you guys would look at just because we have a lot of measures coming through.

And we thought having you look at every one of them would be a little bit too much so we said let's do complex measures and we defined complex measures by outcome measures so any kind of outcome measure would come to you guys.

Cost and resources measures, efficiency measures if we had any of those; we don't have any of those yet. But should they come in, you would see those and also composite measures.

And I want to go through these next slides fairly quickly but I think it there may be a few – if you have questions, it's OK because this needs to be a conversation today.

But I did want to just give a little bit of context for us today and a little bit of level-setting actually as well. And some of these things come from the various conversations, calls that we've already had and a lot of these email threads that have gone back and forth. We've been pretty robust on our emails and that's been great.

So, first of all, I think we've all understood and realized that terminology methods and even philosophy varies greatly just in general but also among our group and we can really see a difference by discipline and also experience.

So, that was probably our first big learning that really has helped us drive today's agenda as well. A couple of things in terms of the desire for a glossary of terms, that is in process. Again, that's a lot of what today is about. And I have in parenthesis here it's part of the toolkit.

Now, this toolkit is in quotes and basically the idea there is that we eventually or maybe not so much eventually, very quickly want to start documenting our thoughts, our definitions, all kinds of things like various methods and approaches that you think are the right ones – excuse me.

What kind of statistics should we use? What are the pros and cons, all these kinds of things that it'd be really great to get on paper. Right now, we're calling the "toolkit." In the past, we talked about white paper, same thing, right? What we call it right now probably doesn't matter too much.

And the other thing just again as level-setting is what we come up with is going to be of interest to a lot of different people -- to you guys because you have to evaluate measures that come across your desk; to developers who bring their measures to NQF would be interested in this.

Standing Committees who want to understand there's methodologists, as you know plenty of methodologists and others on our Standing Committees who would be interested in these kinds of things; NQF staff really want to know these things because we do preliminary evaluations of the measures that you don't see so we need to know about these things; and the public at large just the whole measurement enterprise I think can learn from what we're going to do.

What that means is that we may have different ways of putting out information. There might be a white paper. There might be published papers. We might have a wiki, who knows how it'll end. It may look different but we will be working on that.

So, kind of along those lines, several of our emails and conversations we've talked about this idea of threshold values. Many of you want to be able to get there so we are certainly going to be talking about that not too much today; I think it might be a little bit out of reach today.

But, of course, if we can come up with threshold values, it would be part of our toolkit. We have to figure out what are the pros and cons of having those, what things will we figure out thresholds for and what kind of information do we need to even get there so, all kinds of things that we'll be talking about in the future.

So, with all that said, today we're going to try to stay out of the weeds to the extent possible. And I know I'm talking to a roomful of methodologists. I know we're not going to be completely out of the weeds but to the extent that we can, we won't get into too much statistics that sort of thing.

What we do have though is a parking lot. And the parking lot is over here, with these big post-it notes. They're empty right now. We could have gone ahead and put some things on there, but we're not going to. But I think probably as we have our discussions today, there's going to be many things that are going to come up.

And we're not going to be able to get today, but we're going to write it down. And we will get to all of these things eventually. So, you can help us out today. If something comes up and you thing, oh, we need to make sure that gets on the board, please help us.

We'll be trying to pay attention as well. But just let us know if you think something wants to be on the parking lot. Also, if you just – if you feel like it at the breaks and you come up with something, feel free to write your own thing on the parking lot if you want to that's fine.

A couple of other things, we all know that no health care performance measure is perfect. So, we're never going to see a perfect measure come in the door. There's always going to be some weaknesses. But what we really need to figure out collectively is what's good enough for NQF endorsement.

And when we say NQF endorsement, I want to make sure that everybody understands that when we say a measure is endorsed by NQF, what we are saying is that we believe that it is suitable for both internal QI efforts as well as accountability applications.

And when we say accountability applications, that's our umbrella term and that pretty much includes anything at all that is an accountability kind of program. It could be a certification. It could be things like pay for reporting. It can be P-for-P or public reporting, so all of those things.

So when you're looking at measures, you have to ask yourself, you may not know that it's being used for a particular use or you might know that it is. But you have to assume that it could very well be tomorrow put in any of these kinds of programs. And with that, (Mike).

Michael Stoto: Maybe you want to spend some time to discuss this later, but I want to raise the idea that between the two of those things, they're in conflict with one another because some measures I imagine could be good for accountability and not for quality improvement. Or some things are good as long as you have big enough sample size and so on. So, I'm not sure how we're going to handle this.

Karen Johnson: Yes. It's a bit of a question that I think we'll talk about later as we go, not – maybe not so much later today but on – further on.

But I think you're right. I think generally NQF's position has been if a measure works for accountability it will probably also work on the QI side. It's not necessarily the other way around, right? So it's something that works really well QI may not be appropriate for accountability.

Michael Stoto: But I think it's much more general than that.

Karen Johnson: OK.

Michael Stoto: It might be – the other example is this is reliable only if you have a big enough size.

Karen Johnson: Right.

Michael Stoto: Or that could be a validity issues in some settings but not on others and all sorts of things like that.

Karen Johnson: Right. Right. So let's not lose track of that idea. Now, as a group we're going to try to come to consensus. And so far, I think we've been able to do that, our conversations mostly have been very collegial.

They've been really great and interesting. But when we say consensus at NQF, we don't mean that everything necessarily completely agrees. Sometimes people think of consensus as really nobody is completely happy, but you can live with it.

That's not our official definition of it. But we really just want to have kind of general agreement going forward on things. Also, please realize and this – I know we talked about this when we first brought you together. You guys are going to be making a lot of recommendations today possibly and going forward.

Your recommendations are not binding to NQF. We actually have another government structure. So what we would do if we make recommendations we

would package that and take that to the next kind of level at NQF, so just so you know that.

And I think that's really all I wanted to say. Let me see if anybody had any questions or comments on context and then we'll get into the – our next major portion of our meeting.

Male: We might hand to you or to (David or Dave).

Male: So just coming back to the point (Mike) made about accountability. I just want to check an implicit bias perhaps that I've had and maybe I'm wrong. But I – if someone were to ask me if a measure could only – you can only be certain a measure could do one or the other, be good for accountability versus QI, I would lean toward accountability.

But maybe I'm wrong about that, mainly because I think local decision can be made as to whether one uses an NQF endorsed measure for QI with the local decision. But maybe I'm wrong, I was thinking we were more concerned with accountability.

Male: And I would ask each of us in our own minds given the work we've been through so far, how many endorsed measures we've seen are one or the other or actually meet that criteria of both. We might all have different answers.

Male: And I think that this is an issue that we will continue to hit over and over again, not only today but as we go along further. And I – personally I agree with (Mike) on a number of these, but there are all sorts of subtleties about when is measure a good measure or a useful measure or when it is not, sort of within each of those two major camps.

And there's a lot of nuances there and I do think we have an opportunity to speak in a little more depth and perhaps a little more detail in previous NQF groups that have spoken about this issue. And if we find some messages that as a group we feel that we can get behind I think that'd be a good thing.

Male: Not to belabor the process, but I will just make one quick comment. I think that the three domains here, there is accountability, there is transparency,

which are related and then there is performance improvement or improving population outcomes.

I think they're all three equally important in terms of measures. I think, yes, we need to be accountable, but really at the end of the day, what we're really striving to do is improve patient outcomes. So I think they're all really important.

Karen Johnson: OK. I think now – oh, yes, thank you. You guys are going to have to help me because… yes. We do have people on the phone. Anybody on the phone have any questions or comments before we get into our process discussion?

(Gerald): This is (Gerald). I don't have a question. I just wanted to make sure my line was still open because I got disconnected.

Male: You're good.

Male: We can hear you.

(Gerald): OK. Thanks.

Karen Johnson: OK. Poonam.

Poonam Bal: All right. So, I'm going to start with a quick recap of the past two cycles. I just want to remind everyone that the fall one was a unique experience because the way that the – when we got the award it was too late for us to the three months before the submission, we had to do our work along with the Standing Committee.

And so that was a slightly different experience, a very shortened version of what we experienced in the second cycle, which was the actual process.

Either way, we found hurdles along the way, so we want to talk about those and try to come to a better place, but just a couple of statistics to start off with. In the fall we evaluated eight measures, seven which were new, one which was a maintenance measure.

Five or those were evaluated by the co-chairs, meaning we were not able to come to consensus amongst the three reviewers and they had to go to the co-chairs for review. From that, four were approved by the Methods Panel to move forward. And three of which was after the co-chair review.

And then with that, one measure was overturned by the Standing Committee. I do want to remind everyone that if the measure did not receive a passing score, it did not go to the Standing Committee.

So the only measures that can overturn are if the Methods Panel gave a positive review for the measure. This measure, the Standing Committee found some errors in statistics and that's why they decided that it was not good enough to move forward.

So that's the situation. In the spring, we had a lot more measures and to quite frankly envision a lot more measures coming in the upcoming cycles as many of the projects will be hitting the three-year mark for measures.

But we had 21 measures, 9 that were new. Thirteen of those were evaluated by the co-chairs. If you can look at the percentages here, roughly to same percentage of measures ended up going to the co-chairs.

However, the approval rating was not as favorable. So, in the first cycle, we had about 50 percent of those measures move forward. This round, we only had eight or 38 percent of the measures move forward to Standing Committee review.

At this point, we don't have data on if any of the measures were overturned because the meetings are set to be happening in June. So at that point we'll have a little more data on if the Standing Committee disagreed with the measures that were approved to be move forward. (Christie)?

(Christie Teigland): Yes. Could you just explain the difference between – I mean which ones went to the co-chairs? Were they ones that were approved by the Committee or dissonance there? Which ones moved forward to this?

Poonam Bal: Yes. Yes. So, of the second one – so of the 13 measures that went to the co-chairs only four were approved to go forward.

(Christie Teigland): No, I mean the step before. Which ones go to the co-chairs? That – I don't think you explained that.

Poonam Bal: Oh. Yes. OK. Sure. So if the measures – we did not hit consensus, meaning all three reviewers did not agree that either the measure was low or insufficient or that they agreed that it was all moderate or high, it would go to the co-chairs.

And then they would have to do their own reviews and then they weren't allowed to look at your reviews by making their final decision.

(Christie Teigland): So the ones that didn't move to the co-chairs were just out at the first…

Male: Or in.

Poonam Bal: Or in…

(Christie Teigland): Or in. OK. OK.

Poonam Bal: … clear they were either out or in.

(Christie Teigland): Because there was 100 percent agreement. I understand. OK. Yes.

Poonam Bal: Exactly. (Matt)?

(Matt Austin): Yes. Would we have been notified if one of our measures is not in agreement with our fellow reviewers? Because I never heard back anything so I don't know if I should take that as, are as agreed or if they were just passed along to the co-chairs and we weren't necessarily notified.

Poonam Bal: I think that's good feedback for us that we should be getting back to you to let know what the progress was. It really depends. So, this round and we're going to jump into this a little bit, but I'll jump in to it already, we had a different structure.

If the disagreement was on the type of data provided, meaning people did not agree that there was data element or measure score reliability or validity provided, then we would generally email you back and see if we can get agreement through email and then initiate what we've been calling resolution calls in order to see if we can get consensus on those items.

And then based on those discussions on the final rating, then it went to the co-chairs. However, if it was not a disagreement on the data provided but actually the quality of the data provided then it would just be sent to the co-chairs directly without any sort of resolution call or emails or anything of that sort. So it varied based on where the disagreement was.

(Christie)?

(Christie Teigland): Yes. It will be useful to have another row to show like the total number of – out of the 21 measures and how many were ultimately went for – were approved.

Female: Yes, so it…

(Christie Teigland): Right. We don't have that from this data.

Female: OK.

(Christie Teigland): Yes.

Female: (Eugene)?

(Eugene Nuccio): Well, I don't know what you said about eight.

Female: Yes.

Male: Or it would be seven. Yes. In the fall the answer would be seven were finally approved because we…

Male: I think…

Male: I'm sorry, three.

Male: Three, yes.

Male: Three. It would be four minus one. And then the spring it'll be eight minus whatever TBD is.

Male: Right.

(Christie Teigland): I thought you said that – that from the ones that went forward to the co-chairs there was disagreement, but the ones that didn't go forward there could – they could have been approved or not approved. So it, couldn't there hae been…

Male: Oh, I see. Yes.

(Christie Teigland): Yes. So I don't think that's correct.

Male: Yes, that's right. Yes.

(Christie Teigland): Yes.

Male: Well, at least let's make sure we interpret it. My understanding would be if the reviewers agreed that something was OK that would not have gone to the co-chairs but it's in the eight. Would that be true?

Poonam Bal: Yes. So it would have gone directly to the panel if everyone was in agreement. And unfortunately, we have the data on our own notes, but not on here, so four of them went through with no co-chair review.

Male: … getting into the numbers. This may be the weeds. But – well, but I think your point is that we know that three were approved, but there's another four – I'm sorry, there's another three, that is eight minus five that might have been approved or not. So, it's going to be 3 plus 0, 1, 2 or 3. And that's the number you're looking for.

Female: That's the number.

Male: Yes. All right.

Poonam Bal: It's great feedback.

Male: And so once we get – and once we get to that number, are all of the developers notified of the action of our committee at this point one way or the other? Oh, OK. Thank you.

Poonam Bal: That's correct.

(Matt Austin): Hey, this is (Matt). I just wanted to endorse that recommendation that the panel members be kept in the loop about resolution by co-chairs. I was part of a review group that had lack of consensus. And I think our measures may have gone to the co-chairs.

And I think it'd be helpful to understand sort of what the decision was and rationale, just again as learning and to keep us sort of plugged in. And I think this is going to be a theme that comes up today, is around communication.

(Dave Cella): Well, there, you're next, but I'll just comment on that that this is Dave. I would even add that maybe there is a way – it's just like we do grant – those of you who do grant reviews, you'll have three reviewers. They'll give a score.

And people are allowed to change those scores, so I think if there's maybe some way to have the co-chair review and then circle that back to the reviewers, because the minority reviewer who may be an approver or a disapprover may say, "You know? I think you're right on second thought." And then we've got real consensus or they may push back. And that process could be useful.

Male: Yes. I second that. I wonder if we could have like a clinical dashboard maybe within a SharePoint. So for each time, for each measure you could post the reviews that have been done. Basically, you do your review first, so you don't get to see other people's…oh, I'm sorry. Sure.

Yes, sorry. I wonder if we could create a dashboard within a SharePoint, so that for each measure after you've completed your review, you get to access the reviews that other people have made and the reviews that the co-chairs have made as well when that's necessary.

And that would allow – that would give you some structure to sort of see what other people have done and at the same time if it has to be an itinerary process you could see what, again, what other people – what other folks have done.

Male: Two points, one is back to adding the line, what we really need is kind of a flow chart that showed which things got decided when and where.

But the other thing is that I agree with the point about the dashboard and so on, but I think it would also be useful to maybe find some interesting cases where there was disagreement and share it not just with the people who are involved in reviewing it but for others. Not to the point – not so we're second-guessing, but so that we learned about where the points of trouble are.

Female: And just FYI, our CSAC, our governing body has actually asked for things, especially want a little bit more detail on the things that you guys weren't agreeing on, so we are going to be doing that work anyway.

But I think you're right. The other thing that we're doing as staff is kind of mentally and physically writing notes so that it comes up later in our monthly call, but I think we can be even more systematic.

Poonam Bal: Were there any more questions on the phone? Oh, I'm sorry, (Ron)?

(Ronald Walters): Thank you to the co-chairs first of all. Thank you for the data too. I realized this is only two data points.

But what I interpret from two data points is that we started out a little bit on the conservative side and then we may if you would believe there is statistical significance there, we may have gotten a little tougher.

And I suspect that trend is going to continue for a while we achieve our goal of improving what comes to us. So sometime over the next couple of years, I would hope that perhaps even though we look at very complex measures, that those numbers swing a little bit the other direction and maybe someday we'd go out of business, but I read into two data points whatever you want to read into.

(Joseph Kunisch): And this is (Joe). Just a quick comment, I think it would be helpful also to know what types of measures we were struggling on. Were they claims based? Were they outcomes?

Were they the electronic clinical quality measures? Maybe some of them are easier to get agreement on and some more are more challenged in like the composite measures and what areas in those measures make it difficult.

Poonam Bal: Perfect. Thank you for all that feedback. Obviously, we need to be better with our communication and we're taking very thorough notes and we'll definitely try to get to more details about exactly where we're – more about the numbers of what's failing, what's getting – consensus is not reached and so on. Thank you.

OK. So just a reminder of the process, some of you have already mentioned but I do want to go into it for people on the phone that may be joining in for the first time. So, one, a minimum of three panel members will independently value each measure.

The assignments are based on expertise, availability, need for inclusion and other assigned measures. And then we do provide a standard evaluation form that mirrors the rating algorithms.

The majority recommendation from these three evaluations will serve as the overall assessment of reliability and validity. Again, this is not a final decision. It does still have to go to the Standing Committee and they can overturn any positive rating.

We do want to emphasize that if the measure goes down it won't go to the Standing Committee. It goes directly back to the developer. But if we vote positively then it will move forward.

And we are estimating the workload and this may change is about 15 to 20 measures per year, broken up amongst the two cycles. And I think we probably haven't seen that these past two cycles because it's the just the beginning, but we will definitely see higher numbers going forward. Just as a fair warning.

All right. So then going back to the process, if there is disagreement in the ratings between the three reviewers, the panel co-chairs will then evaluate the measures and determine the overall recommendation.

We have found that this takes a lot of NQF staff time trying to curate all the documents and try to figure out where we're getting disagreement. And it's definitely more of a burden on the co-chairs than we initially thought.

We had thought, only a handful of measures will be going to the co-chairs, but we're finding that a lot of them are actually going. And then once that is all done and the co-chairs come to consensus, staff do compile the ratings evaluations and the comments on the reliability and provided to the NQF Standing Committees.

Again, the work of this committee is to inform the Standing Committee's endorsement and the Standing Committee can overturn that decision, so now, sorry, learning how to use this clicker thing, so lessons learned and some course corrections that have already happened.

One, we've definitely realized more information is needed for your evaluation. The panel has their work after the intent to submit but before the actual submission. So when we initially envisioned it, we didn't realize how much information we actually needed at that intent to submit time for reliability and validity.

So some things that we've added already, so for maintenance measures, on seconds like, oh, we did now start providing a summary of the last evaluation. So you would know what the Standing Committee had previously said about the measure and what their feelings were about it.

We now provide the feasibility scorecard for eMeasures or eCQMs. And then going forward, we will provide full measure specifications. I know a lot of people felt uncomfortable making that – deciding that first question about, are the specifications appropriate, because you didn't have all of them. So, with the next cycle, we will be incorporating all of those.

And then just a staff perception, submissions often do not provide enough detail about the methods that they did. So it's not always easy for someone who may not be an expert in the work that they're doing to completely understand, OK, they were trying to do this method or they kind of skipped over this in their description, but this is what – how they got to the result.

And then the second thing that we've noticed is difficulty with the evaluation form. So we've had many conversations about the form.

So the MP – sorry, MP members have had trouble with the form, mainly around there has been some issues with kind of the skip logic and is that really how you want to do it and the order that things go into and why do we go in that order.

And then there is a desire by many for – to people to write more not less. When we initially envisioned the form, the goal was to make it easy as possible on you all who are reviewing the measure and keep – make it more palpable.

Unfortunately, we found that unintentionally we've made it – haven't encouraged as much reflection as many people want.

We've heard feedback that there's not an opportunity to provide as much feedback if the rating is positive. Often if you have a negative rating then you're encouraged to write down why. Let's see here. And then the staffs do use the same form.

Female: Yes.

Poonam Bal: Just give me one second, (Mike) and I'll get to your question. Thank you.

And the staffs have to use the same form for the noncomplex measure and they are not – they canceled this one either, so you're not alone. And then often there's not enough detail in the submission and the form for us to always understand how we got to the end result. So with that, (Mike), what was your question?

Michael Stoto: So I'm one of the people that have trouble with the form. And I found myself spending more time kind of struggling with how to follow the logic, the tree than I was making scientific judgments. I don't know whether others have that same experience, but I'm not sure how to address that.

Male: I think that is sort of part of this next hour's discussion we're going to have, because there are some different options that will be presented to us that range anywhere from minor tweaks in the form we have to moving to a much more open-ended thing but that, of course, may have its own downsides about if all of us just write something relatively free form, how do the people in the far-end of it then put that together. But this is the next hour's discussion.

Poonam Bal: Yes. Exactly. So we will get to that question for sure. Any other comments before I move forward? OK. So then next it's – some things that we've learned from the evaluation process, so completely independent evaluations are not yet working as desired.

So in the second cycle, we did allow informal discussions between the evaluators, so we encouraged you guys if you found issues or had concerns to reach out to your fellow reviewers and talk to them through email or phone but still do separate evaluations, have your own one.

We also found there was a need for extensive review by NQF staff to ensure consistency. So, again, we've also noticed the issue with the flow some things are skipped that shouldn't have been. Some things should have been skipped and so on. And so, oftentimes that involved incorporating phone calls to just jump to the end point we found in the first cycle that emails weren't as clear.

We were taking so much time to write up the email and then you're taking time to respond. It was just easier to just jump on the phone and talk through it. So also, additional guidance needed.

I think that's very clear that as Karen mentioned earlier, there is not complete consensus on even the basic definitions or how the criteria should be applied. And our goal is starting with today and moving forward that we can build that guidance for you.

But there's other things like for risk-adjusted measures, inclusion or not of certain factors in the risk-adjustment approach should not be a reason for rejecting a measure. So we wanted to provide that additional guidance of, well, yes, it's important, but that shouldn't be the only factor bringing a measure down.

For all measures, for guidance, for incomplete or ambiguous specifications are grounds for rejecting a measure, but remember there's an option to get clarification, although we want this done early on.

So, again, if you're finding things are confusing, reach out to staff, reach out to your co-reviewers and we can try to see if we can get the developers for that clarification if we can't do it ourselves.

And this toolkit is going to be a major theme throughout the whole day. So hopefully, some of these issues will be resolved with that toolkit. OK. So, I think it's pretty clear. We still have two key challenges with the process that won't necessarily be to fix with more guidance or more communication.

One is a lack of consensus between panel members. And that can be three-folds. It can be as I mentioned earlier a lack of disagreement on the actual data provided. We don't agree that they provided us data element reliability or and so on.

And then it also can be disagreement on the final rating, so maybe two reviewers think, "OK, great, this measure is good enough or great" and then one reviewer is like, "No, not at all. This should not move forward."

And then also disagreement with the co-chairs, right now, the co-chairs do independent reviews. They do see the three initial reviewers' review but they do their own.

And so, there can be disagreement all three levels and trying to resolve that it's quite lengthy and much more lengthy than we thought it was going to be. So it definitely puts an excessive burden on the members of this panel, the co-chairs, and staff. It also causes major delays in the workflow and confusion regarding timelines.

It's difficult for us to tell project teams or developers exactly when they're going to get their results of their measured view because pretty much every measure is on its own timeline depending on where we don't hit consensus.

And then there's definitely uncertainties in the handoffs to the Standing Committee. We give five reviews. We don't say who did what review.

And there's not really clarification for the Standing Committee on, OK, this is the co-chair final decision other than the rating. There's no summary of here's where the disagreements were and here's where we came to consensus in the end. So that's something we definitely want to work on.

And then as (Mike) mentioned, definitely dissatisfaction with the evaluation form, it's something that we need to continue to work on and improve as we go along.

(David Behrens): Just quickly, any other questions or comment.

Male: So I think it's true that permission do not provide enough information and I think that may contribute to the…

(David Behrens): Speak in the mike please.

Male: So I agree that some submissions do not provide enough information and I think that may contribute to the lack of consensus on the panel because now you have to make your own judgment try to kind of how do I interpret limited information. So, I mean, I think we should ask submissions to provide more detailed information so that may solve some of the problems.

Male: One other challenge that I experienced in the first go-round that maybe others have experienced was perhaps in the numbering system that NQF uses. So, for example, in some cases an assessment instrument has a number and from that instrument numerous outcome metrics are calculated.

If PROMIS, for example, have one number, OK, and you have numerous things, it makes no sense to figure out what it is that I'm assessing in terms of

the outcome measure associated with this because there are so many sub-elements.

So perhaps NQF can revisit some of its own numbering system to make sure that we're being asked the question that's specific enough so that we can give quality feedback.

Poonam Bal: And just real quickly going back to CQ's comment, I think we probably agree and we're going to be trying very hard to figure out the types of things that submitters should be putting on the forms, that sort of thing and communicating back to them, so I think that will improve over time. Just to go back and just to make sure everybody understands if we continue on with kind of the similar process.

The idea, we don't have a lot of time to get more information, but there is a little bit of time. So if you look at it real quickly, you get your measure in your mail and, "Oh, gee, I need to do this evaluation," if something right off the bat is, you know that there's not enough information, we have a little bit of time that we can go back and try to get that from the developer but it's a very short window.

Otherwise, if we kind of missed that window, you really will have to land on insufficient and at that point it has to go to co-chair and then probably back to the developer. I think, again, over time people will get better at adding stuff, but that's kind of the process.

Male: Might that be, I'm sure some reviewers might be able to take a quick look and make sure there's enough information. But what about maybe an – and maybe you already do this in staff reviews of information.

Karen Johnson: We actually do. Before we send this off to you, we look at them and we do what we call a completeness check and it's a little bit more than a completeness check.

We look to see if certain things are answered and we try to look and see if there seems like there's enough, that it's answered, it's responsive, it's not just something written down but it's actually responsive to what we asked.

And we sometimes write quite long emails back to the submitters and list out the things that we think they need to add. We give them a couple of days to bring those back to us. Most of the time people do, sometimes they don't take our advice and they don't always put in the things that we suggest that they do. So we do try that and with variable results.

Female: Hey, Karen? Appreciate that. Is NQF offering workshops or educational sessions on this issue in particular? Or is there something we can do to help educate the submitter?

Karen Johnson: Yes and no. We have a yearly and it's kind of an annual what we call developer workshop. So in the past couple of years, we've talked mostly about validity as we talk about that.

We have a kind of dusty document that we call what good looks like where we try to show those kinds of things. And we actually do a quite a bit of one-on-one, we call it TA or technical assistance.

But there's often quite a bit of back and forth between us and people who are submitting to try to help them. But I think once we get a little more clarity from you on really specifics of what you'd like to see, we'll be able to give even better advice and make it more systematic than it's been in the past, yes.

Female: Yes. If I could just do a quick follow-up, I think I wonder if we could get measure submitters who are really good at this, I wonder if they'd be willing to – I know some of it becomes proprietary, but if they'd be willing or if we could call out some that are best-practice because much of it is on the website.

And if we could just, I don't think a star rating, but if we could maybe point out what good is by case study versus describing it.

Karen Johnson: Yes. We've thought about it and most of the time it's kind of funny. We have many developers or submitters who do really, really great submissions, but sometimes they're lacking on one little thing.

So it's a little hard to give a full one, it's pristine on every little thing, they might do really well on reliability and not so well on the validity side. So that's why we haven't actually had a couple of full exemplars.

(David Behrens): Hey, we have (Mike, Lacy) and (John) in the room, but there may be someone trying to get on the phone. Someone on the phone have questions?

(Sam Simon): Yes. Hi. Thank you. This is (Sam). I was wondering going back to I think it was (Zhenqiu's) comment about the material and the quality of it and the completeness of it and I think that went to a comment like Karen said that NQF staff will often do a review.

But there are some times and this happened in the group I was in where material was submitted and it was like complete from the sense that it was there, but it wasn't appropriate.

And everyone in the committee kind of agreed when we finally did get on the phone that what was submitted wasn't appropriate for the type of testing that they did,that they were trying to address.

And I almost feel like one potential process improvement and I think this is implied later on, is to have a call early in the process, not so much relying on email communication, but having a call at the outset to make sure, OK, is the data that we need there and is it appropriate.

And that's something that the committee, the committee of three, could probably do fairly quickly depending on the measure. But that may be one way to resolve or solve a problem where you're realizing too late in the process that you don't have the right data to assess reliability or validity.

(David Behrens): Just for feedback to (Sam), you have a few nods around the table in response to your suggestion that you can't see.

(Sam Simon): Appreciate it.

(Eugene Nuccio): Two points, one is that I like (Susan's) idea a lot. I want to point out that good has two meanings in this context and I think that what good looks like means

it has a good – a measure meets the standards that NQF lays out. I think another sense that we ought to focus on is what does a good application look like?

What is the completeness, appropriateness and so on? The other point is I'm wondering whether we can in addition to having early calls is actually capitalize more on the staff review.

So one idea might be to say, OK, the staff reviewed this and we see that so and so is here and so and so is not there, and the real issue is is this particular thing appropriate or do you think that this one meets the standard, really focus our attention on the judgment that we ought to make and take less focus on less following the kind of concrete steps in the algorithm.

Male: … the order, but, (Paul), we all did the disclosures of interest earlier. Yes.

Female: Mike please.

Male: (inaudible) University and unfortunately I have no conflicting interests or disclosures.

Lacy Fabian: Lacy Fabian. So one of the things I wanted to ask also for the committee members, too, is in terms of timeline, because some of the suggestions and things that people have been talking about with process are kind of things that might be more immediately impactful but maybe not the types of changes we would like to see for the long term.

Like, for example, in the long term one of the things that I feel like could get at both of these process issues, a lack of consensus and the evaluation form is something that's a much more clear cross-walk between the assessment form that the measure developers are submitting and the evaluation that we are performing.

So it is much easier to distinguish on each side, so the developers submitting their measure information can see what we are going to be evaluating based on what is submitted in that section.

And then we can quickly look to see, oh, they do not appear to have included these key data points that we need in order to move forward with our evaluation. And I would imagine something like that would be a more significant change and something further out.

So I just wondered if there were also any parameters to share for us as we're thinking about it or maybe ensure, we're all pretty familiar with the current state and ideal state and future state and all that kind of stuff.

But if there's something we're looking at trying to implement in the next few months so we have more of a bucket of we just need to do this right now versus where we're trying to get big picture to really I think move the needle to borrow another one of those phrases we're all very familiar with for what we're looking at and expecting from the developers and our evaluations.

(David Behrens): Just a quick response, at the end of our day today we do talk about next steps and we sort of come back on this issue of a white paper, maybe some of those suggestions where you can find some traction in there where we not only are writing sort of conceptually about what does NQF believe about reliability and validity of measures but also (inaudible) forward including instead of a (inaudible) submit and then what we're asked to review, and are they hitting the same issues the same.

I think we'll have a chance to hit that in a bit more detail later in the day when we talk about sort of the work products that we're going to carry out.

Female: Yes. I think that's a great explanation. Today, this session is really I think the short-term, get the logistics out of the way and hopefully that makes the process better in the short term until we can really dig deep into all the conceptual work and how best to really make this work a little better. And I think that's what really that last session about the toolkit is going to talk about.

We're not actually going to jump into the toolkit during that session. It's more about what do you think we need in that toolkit in order to give developers and you what you need. So I think that comment will come in a lot at that time.

(David Behrens): OK. (John, Jennifer, Bijan), in that order.

(John Bott): So earlier on in slide 15 I know that going forward we have to get the full measure specifications. So do you mean the full NQF endorsement form that the measure steward submitted? So you're talking about – you're shaking your head no and so you mean numerator, denominator, et cetera.

So I would make a case where I would like to see, and I went out of the way and for the first round I did review the full NQF evaluation form. I thought I needed the context to understand what was going on here.

I think it's less than optimal to be evaluating this measure even though it's just a scientific acceptability with just the testing format and because you're trying to get your head around what are they trying to do here, and I think sometimes that's insufficient information.

And although it seems like you're saying will not get the full packet of the NQF evaluation form that they submitted, I'll just observe that when I did review the full packet before against the testing form, there was a great deal of inconsistencies that the measure steward had noted.

It got very unclear what this was, what the unit of analysis was, et cetera, et cetera. You can tell different people are filling out different parts of different forms.

So I think a general suggestion, whatever the material we're getting at NQF not only look for the completeness of the material but the consistency of the material, that we're hopefully not getting mixed messages across the material that we do get.

(Jennifer Perloff): Just the form alignment question was a great point. I just want to second that motion. The thing I wanted to point out about the consensus building, I thought about this in terms of reviewing a journal article and I sat down and said, wait, this is a really different experience because – and sometimes we all need to be trained to have inter-rater reliability and, in fact, we're sort of training ourselves through this process.

So when I sit down to do the review, it really is very different than doing a review of a journal article because I need criteria in my head. So that glossary and definition document is really an alignment concept. It's more than just kind of definition.

(Bijan Borah): So this part should be done fast). Oh, it sounds like we are kind of moving towards more like a nice review format I think. I don't know why we don't go that route because it seems that might potentially reduce the burden on the co-chairs.

I mean, for the ones that are (inaudible) for which everyone agreed, then that would just move them and you don't (inaudible), probably you don't have to interfere.

But for the ones that there were disagreements, then those same people or reviewers can actually comment or like playing enough to a meeting like full meeting as to why we disagree on.

And that might – then you guys or we all can basically come to a conclusion, OK, maybe even the person who disagreed may come too. I mean, I don't know why. I mean, I get – what I'm asking is why are we not going that route.

(David Behrens): I may have to defer to my more senior co-chair. I'm new to this role, but I think the co-chairs might be highly motivated to move to a process where they have tie-breaker role to do because it takes time.

Male: Well, (Ron's) predictions are correct, it's just a matter of time and all of our jobs will be easier so we're hoping for your accuracy on that. Something I'm thinking you probably don't realize is how we did the reviews.

And I opted so Karen and I, were given the submissions where there was not consensus. And I opted to do my review independently of Karen, so I did see where the reviewers were disagreeing on and I knew that going in, but I didn't know what Karen was going to say.

So that to me was a good process, because then we had sort of two independent judges and we agreed with one or the other side we were pretty

confident. But then there were times when we didn't agree and those would be particularly useful times to, I mean, I don't think we can do it always, but to bring it back to the original reviewers, that doesn't lighten the workload.

But I do think it moves more toward that NIH type process where reviewers are allowed to reconsider after hearing from other reviewers. You do your independent review and then you hear from other reviewers. I personally agree that that's a good direction to go even though these are not NIH grants, but it's a good process well worn by time.

Poonam Bal: And with that, I think it'd be a great transition to the next part where we're trying to do that. But, (Matt), did you want to say something before we got there?

(Matt Austin): All I was going to say was I guess I would just sort of second this idea of maybe getting the reviewers together first to talk through where we agree and disagree and why we disagree.

Maybe amongst ourselves we can reach to a resolution before advancing it to the co-chairs. So I would see the co-chairs as maybe as a roll up if we as a threesome can't figure out a solution.

Poonam Bal: That is one of the solutions we proposed so…

(Matt Austin): Yes.

Poonam Bal: So with that, I think a lot of these topics have already been hit, but we're proposing two options. And, again, within the options there's leeway and different things that we can do and this is to improve the workflow.

So the first option would be to keep the process as is but with relatively minor changes. So keep the three separate evaluations, we'd have early resolutions of issue between evaluators, OK, going straight to calls instead of emailing.

That could be once we realize a consensus is not reached, we can have pre-planned calls that we cancel if we don't have any issues, so on. So kind of

jumping to that resolution as soon as possible and hopefully that will relieve some of the measures that are going to the co-chairs.

And then also making a simpler process for the co-chair review, so instead of them doing their own independent review and then trying to get them to come to consensus, it would be they will review your recommendations and have a phone call, discuss them and then come to consensus that way of, oh, this is where the disagreement is, this is the way we're leaning towards that.

And then, again, we could always have the option of involving you in those calls or giving you a summary after the fact or anything of that sort. But that's option number one, and obviously these are things that we've come up with. But if you have other ideas for keeping process as is but slightly altered, this would be a time for that.

And then the second option is a shift to a completely group discussion decision, so no more independent reviews. Kind of similar to the Standing Committee, either we would have the full panel or a subgroup of the panel discuss the measures. So, again, the group does all the measures or the subset does a subset, basically the difference between in-person versus webinar.

And then all recommendations are made at the meeting, so you would vote and you would decide (inaudible) again not binding votes, still recommendations to the Standing Committee but the recommendation would be made at the meeting. And then instead of you doing your own evaluation, staff would summarize the discussion and that's what would be provided to the Standing Committee.

So those are some of the three different options, there are pros and cons to both options. Obviously, with option two you'll have more easy to manage workflow. You'll know exactly when you're needed.

It'll be a little less stress on you. Your still have to review the measures and do that part, but hopefully it'd be a little better and less back and forth and everyone can have those open discussions and help us come to it obviously.

With the other one, you still get that independent review. So you don't want to lose independent thought in group think. And so the first option really allows that to happen and is more similar to what we're doing right now. So with that, Karen, did you want to add anything?

Karen Johnson: I think the only thing I would add is reiterating that there's pros and cons on all of these things. We would ask if you – we'd love to hear from you kind of where you would prefer to go if one of these.

You may end up making an option 1.5 or something like that or a 3.0 or something. So this isn't your only, you're not restricted to this necessarily, but these are kind of very different ways of thinking about it.

One is independent reviews kind of as is with trying to fix things kind of on the backend like we were talking about. The other is going to more like what Standing Committees do and it's a very different flavor from what you've done so far.

(David Behrens): OK. I've got five people indicating desire for comment, but also we can go all sorts of different directions here. Let me just look to (Paul) first because he went up first before we even got fully through this.

So let me see this. Well, I just want to say because at some point I want to do a quick just show of hands on general preferences and maybe before, but, (Paul), you were up earliest.

(Paul Kurlansky): Just two thoughts, one is and I haven't really finalized this in my mind, but the advantage of option one is the independent review, the thing that it actually comes up later in the presentation, but you suggest the possibility of shortening the time for the window from four weeks to two weeks and that might facilitate the process so that might make that.

I'm a little cautious about relying everything on calls because just from a logistical point of view I answer emails at four and five in the morning and call is another dimension. So I'm concerned about getting everybody that needs to be there on the table might actually delay the process more than the emailing back and forth just from a logistical point of view.

The other thing about option two, it could definitely work very nicely and certainly lower the burden on sort of back and forth. But the issue with that is people have to be very committed to come to that meeting and if they don't it will fail.

(David Behrens): And I've got other the four people teed up and I'm sure other will jump in. Just in terms of sheer quick gut reaction before we get into the detailed discussion, quick show of hands, those who would tend to favor option one? Two? OK.

Male: On the phone?

(David Behrens): On the phone? I know this is kind of hard. Can you just quickly say one or two whoever's out there?

Female: Two.

Male: One.

Male: One.

Male: One.

(David Behrens): Well, that's interesting. OK.

Male: Three ones.

(David Behrens): So before we start this in detail, there's kind of a tip to one but it is not universal so let's just keep that in mind as we go through. And let's – suggesting we need to end up there necessarily because there may be very good arguments. I just want to get a sense of which way this was leaning before we started. So at that point, I have Larry.

(Larry): So I'm in the minority.

(David Behrens): But often the minority is right. We just don't know.

(Larry): We don't know.

(David Behrens): I've been in many groups so I found myself in that position. It's OK.

(Larry): So option two is the approach that is used in the Standing Committees. And it's a really good way of, one, achieving consensus and, two, taking advantage of all of the amazing expertise that you have within the group.

When all of us have different areas of expertise and when you sort of divide this process into small groups, I don't think we take advantage of it.

I also think that when you talk about a measure in a group, a lot more of the evaluation, it's a much more robust evaluation, you get a lot more discussion, a lot more insights whereas when you're sort of sitting by yourself and doing this on your own you can't really take advantage of really talking to other people. And even if you reach out to the other two people, it's still a very small group.

The big question I have around option number two is the feasibility. I mean, how often would we get together? I mean, these in-person meetings are fabulous and I don't have a lot of reservations about people preparing for these, because I think really people will do the preparation.

What I worry about is how do you do this in a way that gives timely feedback to the Standing Committees and doesn't have us meeting every two months. And so I think if we could maybe address the feasibility question first, I think that would be really helpful, hoping more people have an opinion on this.

(David Behrens): All right, thanks.

I understand we have (Jack) interested in a comment from on the phone, and then I'll get back to the folks in the room.

Jack Needleman: Great. Thank you. I think my tilt towards option one is precisely over the feasibility. While we're not seeing all the measures as they're coming through, we are a key bottleneck in this process.

We've got a lot of measures to look at and then they get parsed out to different Standing Committees and have a lower volume of work and just in terms of total number of measures they're considering.

I think what we're missing from the process that we see in the Standing Committees if you think about the process there of several people (the rest) take the lead in presenting then there's discussion and then there's voting.

We were losing two things right now, one is we're losing the opportunity for people to learn from one another and to learn not only about the specific measure, but as we've debated within ourselves and seen things around what validity or reliability means in a measure learning from measure to measure.

So I'm sort of in the option 1.5 place, not necessarily for more meetings, but I do think it makes sense after the independent review for the members who are looking at something to talk and try to reach, see if there's an understanding of where the sources of disagreement are and whether there would be consensus and move for that.

And likewise once the chairs get involved, co-chairs get involved, whether that discussion could be expanded. And then if some of the sources of disagreement and tension can be shared with the larger committee so we can learn from measure A as we think about measure B, C or D.

We've been doing that indirectly in the discussions we've been having on the phone monthly with validity and about validity and reliability, where people brought up examples from their things, question is whether we can make that a more systematic process here.

And those I think are opportunities for improvement and learning from one another that are currently missing in the atomistic review we've got.

And I think a lot of the process here was based upon what's turned out to be an erroneous assumption that the application of the criteria would be relatively straightforward, there'd be a high level of consensus among the three and therefore little need for the chairs to get involved, little need, and an

assumption that whatever the three people will put together from the broader group would basically come to the same conclusion.

And I think those two assumptions have been clearly raised and have been challenged by the results to date and we need to think about how to overcome, how to deal with the lack of, less consensus that was anticipated with the three independent reviews.

(David Behrens): Thanks, (Jack). And I think it's important for you to know since you can't be in the room, there are a lot of nods on your statement about learning from each other which I think for which there's support that we've seen.

And also even more nods and smiles about the erroneous assumptions. So I think it's some shared basis from what broad issues we're trying--

Male: I think my comments are similar to both of those. And I think that kind of the principle I would like to think about is it has to do with what's the contribution of this committee and how best can we draw on our expertise and experience and so on.

And I think that would be to try to find a way to focus on the substantive issues does a certain bit of evidence introduced really do make -- does it really make the case that this thing is properly valid and so on not whether or not they have checked all the boxes.

Now, I think that moving is not all the way from one to two, maybe in that direction and haven't get together either in person or an organized way on webinars I think will help with that.

(David Behrens): OK. Bijan, I also have (Karen) on the phone. I'm going to slot (Karen) in there after Bijan.

Bijan Borah: OK. I think I agree with what has been said so far. So one of the things that -- so the reason why I want to (inaudible) because you see how they are portrayed and I don't know how important that is and I think on their behalf, the question is what the developer gets to see, you know? Once we -- so for example, our method is our method is rejected at our committee level.

We just only go to them, right? And then for them to actually improve our committee, so I don't know as to what they get to see in terms of do they get to see our reviews in their entirety or do they get to see summary of what we have said, I mean the reviewers have said. So that is the one important thing.

And the other thing is, again, I think in the second review that we had recently, I mean, we had about 21 sort of measures that we've reviewed. And I think we still want to have option one because you know, you don't really want to serve 10 times. You can't finish like 21 measures on a single day, I don't think.

So for the ones that everyone agrees you can skip the discussion and only the ones or discuss only with the ones that there have been disagreements, and again, we are going back to (inaudible) and I mean that's not what you do I guess.

Male: I'm just thinking the same words, that where we may find some common ground here in the 1.5, 1.6 territory or something like that is in addressing the issue of feasibility, is that having a whole group discussion is feasible if we're only discussing those on which there's some disagreement which is essentially to this date has been a co-chair issue.

But I think what I'm hearing is some support for the idea that we would learn from each other and actually to some extent enjoy the experience of discussing this because we would become smarter people after we have done it and there would be probably a stronger sense of feedback to the developers and Standing Committees if the ultimate message came from the group as opposed to two people who are subset of a group.

But if those that are uniformly rejected are not discussed and those that are uniformly approved are not discussed, at least not in detail and they can be on the agenda just to note that they're there, very much in the end that may address the issue with feasibility and the ability to have a whole group involved as opposed to the co-chairs themselves.

Karen Johnson: And just real quickly to jump in, we didn't give you all of our thinking, but for option two, if we did some kind of group thing, we would definitely ask you to do some sort of pre-evaluation, so that we would know upfront, number one, you've had a chance to look at things and where the areas of disagreement would be.

If we did, getting a little bit to the feasibility, we thought about it just a little bit, at minimum, I think we'd have to have one two-day meeting per cycle, so you'd have to come to this meeting a couple of times a year for a couple of day.

And doing potentially 50 measures in two days would be a huge lift. So we would hope that over time people would generally, we would have fewer that we'd have to look at, but it's a gamble.

The other thing that we didn't mention with option two is if we went that way, we would actually open those meetings up or at least the idea would be to open those up to developers and the public.

So right now, folks see what you've done and what you saw, they do get the full evaluation form that you guys go out, after everything is said and done, we make that available publicly and to the developers if it goes through.

Option two would probably say hey, you guys can be in the room. That may be good because just like in current Standing Committee scenarios, you could ask developers questions, et cetera. So again, just a couple more details of what we were thinking around option two.

(David Behrens): I think we can Karen on the phone wanted to comment or question?

Karen Joynt Maddox: Yes, I guess I'd put out there and I don't want this to sound like too much hyperbole, but I think the consensus that -- and making this as systematic as possible is really critical, because we're talking about billions of dollars in the healthcare system moving around based on measures that people basically see a stamp of approval from NQF and decide that they are good enough to be used to judge performance, to move market share, to steer patients.

And I think that the work it will take to build up a collective sense of where we land on this and that requires learning from each other and it requires building some sort of group think so that the next iteration of this group in two years and four years and six years and eight years continues the same systematic approach I think is so important.

And I think the thing that was most striking to me about getting all the measures to review is just the lack of consistency and through nobody's fault people interpreted the measure developer submissions different, the quality of the measure developer submission is quite different, the length was quite different and the criteria we all use is quite different because none of us knew exactly where we should be landing.

And I don't think that's an acceptable long term place to be. So I would vote for something in which there's a collective mind melding I guess and be that in person or not, I would raise two other small points, which is I totally agree with the modelling around sort of the grant review where you know you get your however many things to review beforehand.

You turn them in. You can see each other's. The ones that are in the top or the bottom or the middle or whatever, the cutoff, gets discussed, other ones don't, I think that's the sort of study section format is a good starting model.

And then finally, I think we also have to think about the subgroup idea. I was so lost on some of the measures that were more survey based or instrument based and that's obviously a place where we have some really deep knowledge on the committee, and I was pretty comfortable with some of the claims based measures that I've worked on before, and I'm sure there are people that feel the opposite, and so there may be some ability to streamline by having sort of say three or four groups within the group where it wouldn't require that everybody be all together for all parts of the conversations.

(David Behrens): OK, thanks, Karen. I've got in the room (Christie, Jack, Lacey and Ron), you put your -- OK, (Ron) is down, all right and (Joe) on the phone, also just as a time check, we're about 15 minutes from break and unfortunately we have even more slides to see that have other options for other issues.

So let's -- I think this has been very productive. I like the track we're on. Let's just keep in mind, we have a few more things to try to fit in before 10:30. (Christie)?

(Christie Teigland): Yes and just a couple quick things, I'm kind of in the middle also. I think we need that independent review because I definitely spend a lot more time seriously thinking about the measure. I mean we all have those tendencies if we're relying on group think to sort of let others do a lot of the work, not me, but I know some people do that.

And so I'm worried about forced consensus, we don't want that I don't think. And then I'm thinking about the burden though on this committee and on the NQF staff and when David was describing earlier, we had a measure and three experts disagreed on whether the measure was good or bad and then (David and Karen) disagreed.

Put it back put the burden on the developers then because that's a lot of people who are not -- right? And why do we have to resolve that, right? I think that's the time you just put it back, let them resolve it and send it back.

And so I think we're probably taking on too much of the developers' work in that and trying to interpret what they were supposed to do or say or think. But I do agree, we need this group discussion across the measures that we don't get to see because we don't want -- we need that consistency.

We don't want some measures to be approved based on some characteristics that we thought certain ways about and others being disapproved because of those same kinds of factors and I don't know if that's happening. They very well may be happening.

(David Behrens): OK, I think we have (Joe) on the phone and then I'm going to get back to (Lacy) in the room.

(Joseph Kunisch): Yes. Just real quick, I agree with most what's been said. You know, when I look at both of these options, I see a lot of coordination and I know we all have day jobs too that we have to work around and I'm concerned about that.

So there definitely would have to be some kind of guidelines and commitment from all the members that on this day of the week you're going to get on this call at this time or if somebody responds email, you should be responding within two days.

And then in the option two, what one really concerned me was the -- it discusses all measures, the full panel discusses all measures.

To really do a good quality review of the measure, I mean it takes at least for me, I spend hours on it, you know doing some research behind it and investigating it, and I would want to make sure that we're still giving the time instead of doing more measures and having to review more measures to limit how many person reviews just to give that quality time to it. That's it.

(David Behrens): Yes, thanks. And I would have to say that I shared some of those concerns when I first saw the phrase full panel reviews all measures.

I thought we'll be here for a week at a time, but I think we already talked about the sort of the NIH approach and we may tweak that a little bit where we really discuss those things on which there's some disagreement.

And again, I would probably (inaudible) to that because I've learned from the discussion and why my view might be different from someone else's view and from two, three other people's view, relative to the little time discussing things for which there was already an agreement.

So this moving forward, we can end up with focusing the discussion on things that I guess it might -- each one discussed might take time, but there may not be that many in any one cycle discussed. Lacy?

Lacy Fabian: Lacy Fabian. So I lean to option one and especially before considering option two, go back to the original process comments about the issues with the form and we're trying to get consensus if that's our issue, I think having that cross walk between the forms to the point earlier about going back to the measured developers, if we could just get what we need from the beginning and can clearly see that, I think that could rectify a lot of our process workflow issues as opposed to having to shift.

Similarly reviewing all measures just seems attainable. The only unintended consequence that initially comes to mind if we only review ones with disagreement, I was a part of some of the disagreement calls and I know that our opinions definitely shifted and potentially also if the measure developers are invited to those calls, I worry that we create a situation where now with the ones we're discussing, those measure developers have greater opportunity to weigh in whereas those measures, we're talking about agreement either way.

So those measures that we said weren't good, those measure developers are not going to get an opportunity to clarify whereas the ones being discussed are.

And I worry that our issues that we are discussing would have like applied in the backhand or retrospectively only because I saw that coming up when we had these initial discussions, because we are still so much in the learning mode that we would create an unfair decision.

(David Behrens): That's a good point. And I think I'm hearing a lot of important considerations about if we try to do something in the 1.5 territory where there is actually a group function, how to make it feasible.

But also as you just pointed out, if we go down a certain path that focuses our discussion on only those in which there's an initial disagreement, that leaves an opportunity for additional clarification and input for those, but not for others that we uniformly rejected and maybe they were uniformly rejected unfairly. So yes.

Male: Just a quick comment about sort of shifting the burden back to the developers when there's disagreement, that's fine except you have to be very clear with them exactly what the disagreements were, because otherwise you're just going to bounce back and forth.

And so you know, I don't think it actually shifts the burden at all, because I think the burden will still be on us to make it extremely clear what the issues were and then to go back to them.

(David Behrens): OK. Larry?

(Larry): I want to just bring back two of the comments that Karen made. I think they're really, really good comments. The first one is what we do really matter.

CMS as we all know is redesigning the entire payment system and wants to link 80 to 90 percent of payments to performance metrics. So our decisions which are sometimes based on what three people think, if we reject the measure, that's a really big deal, OK?

The second thing is judging by how quickly we're moving today. We're not going to come to a consensus on how to evaluate measures, and that's why when you have three people, that's why people never agree.

We, as a committee don't really know how to evaluate these measures. And the only way I think that we're going to get to the point of a fairly consistent approach to measure evaluation is doing these kinds of in-person meetings, OK?

That last comment, feasibility and I brought this up in my first comment, it is really resource intensive to bring a whole bunch of people together. I can totally get it. And the other thing is the idea of reviewing 50 measures, I mean that's just not tenable But that's not how the Standing Committees do it.

The way the Standing Committees do it is basically you have three people, two or three people who are assigned a measure just like right now and you go to the Standing Committee or the Methods Panel and you'd have three people take the lead, would present the measure and then there'd be a discussion.

So all of us would not be reviewing 50 measures, all of us would be reviewing a subset of measures. I mean you'd have some responsibility, have some understanding of all the measures, but you wouldn't need to do an in-depth review of every single measure, OK?

And still 50 measures in two days, I mean Standing Committees typically review three or four measures, sometimes six or seven, at least -- I mean it varies, 50 measures is really hard.

So there would have to be some kind of triage process if in the initial evaluation where those, again, the three lead reviewers say this is a fabulous measure, it is terrific, you don't really need to spend the time with the entire committee.

And vice versa, if you have a measure that's just horrible, I mean, so you might end up with maybe 15, 20 measures to review over a two-day period. And I think that would be doable.

And I don't think it would require that much more time than what we're doing now. But again, it would move the process to one where you have consensus. You have 15, 20 people deciding whether a measure is good or not, realizing the potential impact of that measure.

And the second thing is again, this idea that we as a group could evolve and figure out how we should be evaluating measures so that maybe a couple years down the line, we really have a way of doing this and everybody is really comfortable with it and it's just the gold standard for evaluating measures.

(David Behrens): Thank you. I've got Matt and then I may have to stop because unfortunately we have even more things to be put in front of us to discuss and about seven minutes to do it.

Matt Austin: Yes. So this is Matt Austin. I think my comments will be quick. I would agree with Larry that I don't think there's necessarily a lot of value. If the three initial reviewers all agree that the measure is good, I think we could probably all set that aside and say that's good and let's move forward.

And so maybe the discussions need to be around measures for which there's disagreement or to Lacy's point, measures that we're going to potentially reject. And I'm also wondering if there's maybe an opportunity somewhere between three reviewers and the full group of 25 of us.

I mean maybe we split into A, B and C and there's 8 to 9 of us in each of the subgroups and that 8 to 9 get together and discuss the 5, 10 measures that are either disagreement or rejectable. And then maybe our co-chairs sort of serve as a Supreme Court if we still can't quite reach the decision.

(David Behrens): Thanks. Let's move on because the other things also are worth our attention and don't want to discount them.

Poonam Bal: Yes. Before we jump forward, I just want to answer some questions, a reminder that it's not just in-person option. There is also the sub-group option for consideration and then in your question early, you asked what is public?

So the developer, if the measure goes down, the developer gets all the reviews with all your notes. Whatever you hand to us after a little bit maybe back and forth, that's what they get too. And, I know (Dave) you have another question.

(Dave Cella): Well, just to comment on that, they get all the comments, but they're not tied to a name, so they're anonymized. And this option two would be anonymized. I think that's just something worth noting that your individual review or comments will be tagged to you in the public domain. So just an awareness that that's holding –

Poonam Bal: OK, thank you. So with that we're going to go to our second main discussion, so about the form. So we provided three options here. First is keep the form as is with minor changes as needed, so maybe (inaudible) some of the flow logic and so on.

Option two would essentially allow a free text evaluation. It would be modeled off of what staff use to do before we handed it over to you.

So it would basically include different recommendations on what to include and then "canned questions" to consider, so it still would provide some guidance on how to fill it out and what to do.

And then the third option would be to meet someone in the middle, so much more free text than we currently have, but still have those check boxes or at some point you have to hit certain things and answer those.

And so we did send out examples of the two measures done in the current form and done by staff previously. I'm going to ask (Linda) in a second to bring it up so you can have it to view in case you didn't get the chance to open it earlier, but we can start the discussion while she's getting that up.

(David Behrens): Yes, also just because I found it helpful although we ended up going a slightly different direction, just again, before we get into discussion, any quick show of hands, I know the options have gone away on the screen, I happen to have them here, option one is basically keep the form as is although some clarifications and what not.

Option two is move to free text and option three is somewhere in the middle. Show of hands just quick as we get started, those who might favor option one? A couple. Option two, free text? That leaves everybody somewhere in the middle, that's OK, that's OK.

I just want to know generally where are we going to try to take this. I guess just speaking for myself, I'm certainly one of those who had to be corrected on -- I'm sorry? There is an option three which I guess by subtraction, so it's only about three, four hands, everybody else wants to be option three unless you like four or five.

I certainly had to be corrected on the skip patterns and sometimes I realized that I haven't done it correctly. Otherwise I thought that instructions weren't very good and I know from our earlier comments there were some issues with that.

So I would think that whatever remains of the structured form certainly can be improved, even if it's just in terms of our own learning how to use it and understanding.

I also I think struggled a little bit with some of the answer options to each question where it was either insufficient or then there might be a couple of

graded options, and somehow I found my thinking not matching any of those three options.

I might say well what, there was, I would say or call it sufficient, but then I think maybe Mike or someone pointed out earlier, what was there wasn't even a target. So it was a little hard to judge it as being moderate or strong or something.

So in general, there's a sense of option three. Maybe we could focus our discussion on sort of what is important and really worth having about the structured form that we have, how it can be improved, but then also if there's some desire to be able to write some more text, what would that be and where would it come?

Would it be like a broad free text field either at the beginning or the end just as a summary or do we want more extensive free text option as we move through the form? Those are some of the questions I can imagine. So I see Gene up and I'll try to keep up with this people who want to comment.

(Eugene Nuccio): I have a couple of different questions, or comments. One, we have the cart before the horse because I mean the papers that we received regarding reliability and validity clearly point out that the term reliability is applied differently if you're talking about data element and metric or the outcome, and there may be even differences whether it's electronic or it's a composite or whatever within the outcome world. And the same is true for validity.

The forms that we have don't even closely reflect anything along those criteria. It's supposed to be important criteria that we're supposed to be providing input on, that is does the data element provide reliable information and valid information, then the questions in the form that we're using need to reflect those issues, not some hybrid between that and whatever it is that the developer has provided to NQF, which gets back to Lacy's point that if we have a rubric that we're using or that we develop, that has this nice, I don't know what it's called, dichotomy, trichotomy something, reliability, data element, what's the question we're trying to answer, does it exist and is it -- does it have the sufficient threshold so we could make some sort of judgment

about its adequacy so we can make the comment that Larry wanted to make that CMS is obviously making with value-based purchasing.

OK, I'm going to pay this agency X dollars for an 82 and that agency fewer dollars because they only got an 80, is that real, OK?

So let's think about the reliability for the data element, for the outcome, validity for the data element and the outcome and focus the questions that we're trying to respond to using that rubric, that matrix and ask them the developers to provide the information that we need, so we can not necessarily check a box, but does the developer -- did the developer provide adequate information, so we could make evaluations on each of those elements?

(David Behrens): Let me just try to do it by a quick summary and I'm sort of playing off the cart before the horse comment, usually that's a negative connotation. But what I think you said and I would agree with this that there should be some structure to the form, but the structure should match more closely.

Perhaps the result of the next two blocks of discussion when we get into very much deeper reliability for a while and then after lunch, validity for a while, if we have a clear sense among ourselves about what the key dimensions there are and what the key, say, levels at which those terms are applied, the form can then match that and we may end up the day with a better sense of it, but it may be that we have to think about the changes after we've had a couple of other blocks of discussions. Is that a fair restatement?

Eugenio Nuccio: Yes, that's sort of where I'm going, yes.

(David Behrens): OK, no, that's great. Marybeth?

Marybeth Farquhar: I have to agree with Eugene. I spent most of my time going to the form and rereading what the criteria are and rereading what the question is and what they're asking. I'm trying to match it to what the developer has sent.

And every time I go back to what the criteria are and when I go back to answering the question, it changes the meaning, so we need to have a little bit more structure with regard to this is what the developers -- this is the order

that they should have it in and then also our form should follow that order so that we can be able to make sense of what we're trying to evaluate.

Karen Johnson: And I think as much kind of detail as you can give us would be very helpful because to Gene's point, we actually think the questions are asking these the same that we really want to do. We asked you specifically about did they do data element testing and if so, what is the adequate method and were the results reasonable, trying to end that way.

So in our mind, we actually have asked you the very pertinent questions. It's obvious that it's not quite working, so we all need a little bit more from you to help us understand that disconnect.

(David Behrens): OK, let me just do a quick check on the phone, any one, question or comment? (Mike) then?

Michael Stoto: On that last point about data elements and so on, one of the things that struck me as I've tried to work through this is that these terms like that have different meanings for different kinds of measures.

You know, for measure -- for process measure, did you do so and so? That's one thing. For an outcome measure that's risk adjusted, I think it probably means a whole different thing or at least I don't really understand how to map all of that and I suspect that there are other versions of that as well.

(David Behrens): I think that is actually an interesting way of teeing up this next couple of blocks of discussion because as we develop and enrich our own discussions and ultimately definitions of reliability and validity, we want to make sure that they agree to concrete criteria that presumably are as widely applicable as possible, but also if it turns out as you said that the term takes on a somewhat different meaning or has a somewhat different criteria in this kind of measure versus that kind of measure, if that is so, at least we have to articulate it and then, I don't know how many versions of the form there also might be.

But I find the world easier to see even if there are many different complexities as long as they are clearly articulated. And I don't mind trees with lots of branches as long as I can see each branch and know what branch I'm on. I'd

rather work in that kind of environment myself than where everything is fuzzy and unfocused and I'm not quite sure where I am.

But I would certainly accept as a premise for continued discussion that these things may play out differently with different types of measures built on different types of data elements and it's just a complex world. That's the domain we're in.

Male: I have a quick follow up and I'm referring back to that paper by John Adams that we looked at a couple of months ago. And I think two of the good things about that were one is it was that reliability. He said exactly what he meant, what he assumed that to mean.

But then, but he also said -- and this is the setting that he was working in. It was a certain kind of measure and a certain kind of setting, but what does validity mean in this setting and how do you measure it in this setting, and I think that you really have to be clear about both of those things before you can make -- say something useful.

(David Behrens): OK, so we're probably going to come to the break soon. I just want to maybe suggest something for the parking lot, but also make sure that NQF is OK with this, and that is if and when we clarify our structure for data element and performance measure, reliability and validity that the submission structure should mirror the structure of the review, so that the developers are asked by NQF to submit information that's organized the same way as what we put together.

That would be parking lot because we go to -- we have to figure it out first and then make that request.

Karen Johnson: And just so you know we do have complete freedom to change that form, but we would need to do it fairly soon. So this is something that if we could come to some agreement today of what we're really looking for and a flavor of maybe the flow, then I'll have a chance to be able to make those changes in the submission form so that you would probably be able to see those by the fall.

If not, it might -- even if I make those changes, you might not see them as quickly as you might like, because developers are already filling out forms now for the fall.

(David Behrens): Yes. We should try to stick with the schedule for the break but (Poonam), you've got something in front of us that I know you want us to think about. So go ahead.

Poonam Bal: Yes. So this was sent out to you, but this is on your screen, an example of what the form that we used to fill out looks like. The one we fill out has a lot more comment bubbles about think about this, don't forget to consider this and look here, look there, so we would consider obviously offering that to you as well.

But this is a filled out form for you to kind of understand where we're thinking for a little more free text. I think we're definitely leaning towards the middle ground, but it would be something like this with still some things that you still have to fill out. Yes. Matt.

Matt Austin: Yes, when we talk about free text, is free text so that I can make sort of notes for myself or is it used as part of the evaluation?

So in our current form, it's nice about forcing me into choices, is it easy to say whether we agree or disagree because there's a lot of information and free text, that becomes a little murky in terms of agreement. So I'm just trying to clarify is it really to help sort of make comments and notes sort of for how I was thinking about it or is it sort of in place of?

Poonam Bal: So it's a dual goal. One would be to help you -- help your thinking, help you go through the form, make it a little bit more of a logical flow, so definitely for that note taking.

But I think it's also for the developer and the Standing Committee so they know your logic. I think right now if you're going positive, positive, positive, you're not really providing any feedback on why you think it's good. You just think it's good.

Obviously if you're saying negative, negative, negative, then it's the opposite. Then you're providing some detail about why.

But the goal is to kind of offer equal amount of information to the Standing Committee and the developer it's positive for them. So it's both goals and so maybe you think your notes are for just yourself, and as you know we make everything available to the developer, they'll receive exactly what you write is exactly what they get.

(David Behrens): And just a quick personal response for that, do you want them --? OK, so I thought in my own mind, I was sort of favoring this third option which would seem to be kind of where we premised that I appreciate having the structure and you just said that either in the doing of it but also conceivably as we come together as a group and I can imagine as a developer (inaudible) I like that.

But then in a number of measures that I reviewed myself, I found myself wanting to write something, in explanation of what I just indicated in the check box or even to indicate that I had trouble filling in the check box because somehow what I wanted to say didn't fit the checkbox.

I remember one measure in particular, I thought they'd done a great job establishing that this measure was a plausible measure but I would absolutely not approve it by use by CMS to start throwing millions and billions of dollars around, they just hadn't reached that level of testing.

They had shown in a small sample of instances that something seemed to be reliable and valid but I said there's a leap (inaudible) practices and I say now you're talking about a huge number. You haven't shown me that the performance is going to vary enough that dollars should be hanging in the balance.

But I couldn't find the way in the form to indicate that. I go and I'm basically saying a lot of yes, yes, yes, yes, yes which was leading me to think well I'm say this, that's not what I believe it, it's not right.

So somehow this combination of the ability to go through the structure and check boxes, I think is valuable. But then the ability to say either in summary

or along the way, yes, but, no, but or no but, or here's something else. OK, this is probably a good time, we're a little bit over.

(Marybeth Farquhar): Do the measure developers have any idea about what the definitions are with regard to some of the data elements and what they are supposed to be putting in these boxes?

It might be nice if they had some kind of walk through or an orientation if they're brand new. I know that a lot of them are not. And the other thing I found is that they're inconsistent, a lot of the developers in what they put.

I had one where it said both reliability testing for the score and for the element were there, but then when you go through it, they checked that it wasn't there. So that's kind of some of the issues that come up and that's where the confusion I think lies.

Karen Johnson: And just a real quick answer Marybeth, the different developers, especially the newer ones, sometimes the newer ones do a really fantastic job because, they in fact would come to us and we've helped them a lot. Other times it's not quite so much with it.

And then another thing that we've -- and this is going to sound awful, but sometimes you don't necessarily want to trust the check boxes that they check. We try to catch that at that completeness check piece if we do, but we don't have time to go through everything, so we don't catch everything.

And sometimes I think you're right, just as definitions and what have you are a little bit murky I think to the panel, I think that's probably true across developers today, so.

(David Behrens): We're going to take a 10-minute break on the phone and here.

(BREAK)

(David Behrens): All right we're going to get started. Well four teams, not bad. It’s close to 10, it rounds down to 10. Sufficient, coefficient. All right. So, we're going to move to the reliability session. And Karen is going to walk through the slides,

and I'll track people that have things to say. Is it OK with you Karen, if people interrupt along the way, right?

(Karen): Absolutely.

Male: OK. So, there's about seven or eight slides by my count. And just to give you a sense of what we need to cover between now and the public comment, which is at 12:30. So that's good. An hour and a half -- we should be able to fix the reliability problems in an hour and a half, don't you think? And stay out of the weeds.

Karen Johnson: OK.

Male: That's the big challenge. All right, take it away, Karen.

Karen Johnson: OK. Again, some of this basic stuff on the slides is a nod to a much of the really rich discussion that we've had already in our call and also in our emails, back and forth.

So, I actually did want to hit a couple things, you may wonder why in the world NQF says things and uses the terminology that we do. First of all, we try very much to use the term health care performance measure, it's an umbrella term.

Often, sometimes we'll say quality measure or what, that kind of thing, but when we say health care performance measure, the idea there is that you might be talking about quality, you might be talking the cost, you might be talking about access, this is our umbrella term.

Now that said, we also understand true performance is unknown, we only can observe. So, that was kind of a big thread that people talked about a little bit with reliability, so I just want to make sure that we're at least we have in NQF do understand that. We also know but it's worth mentioning here.

And I think we'll probably talk about it a little bit more under validity, but this idea of performance reflects more than just quality or access. There is -- there's other things in there, when you see different values for various

providers for measures. It's -- there is more in there than just their quality of care, so there's other things in there.

Again, we're going to push that off for now and maybe talk a little bit more about that under validity. But I'll just ask your willingness not to worry too much about our label, unless you just really, really have a problem with it, in which case we can talk about it.

But the other thing is, you'll hear us mostly talking about providers. That is another umbrella term, sometimes people use the, see the word provider and they're thinking about an individual doc.

That's not necessarily what we're thinking of, when we say provider, it's just some entity that's providing care, and we use that very generically. You guys don't have to necessarily, but I just want to make sure you guys know what we're talking about when we say those words.

Data elements. This has come up quite a bit, and you guys are right, we actually have -- I don't think formally defined what we mean by data element. We've kind of tiptoed around it. In our trainings that we do, we often talk about data elements as the building blocks of the measure.

You know, you can also think about them as the variables that you use that go into a measure. So, obvious ones for the more simple measures are things like the diagnosis codes your (ICD10) codes, the dates, somebody's sex. Those kind of things, all the things that go into building a measure, those are the data elements.

It gets a little bit trickier when you talk about surveys or in other kinds of instruments. What are the data elements there? We consider the questions or the items as the data elements. So that's what we're thinking of. So, that's our thinking there. Oh, I'm sorry, Mike, I didn't see you.

Michael Stoto: No, I think is very helpful to do. I just -- but the ones that I'm -- I think you maybe just now getting to the one I was going to ask about is, when you have something where there's kind of a questionnaire in there, maybe 10 questions, and then some calculations made they're summed up or something that

created a score. And that this -- I guess that's the score and their answer to your questions is data elements?

Male: I think the score would be the data element.

Michael Stoto: I mean, I know.

Male: And then there is the reliability, you mean, you can tag to it, but -- right?

Michael Stoto: That we have data element and score, two different bullets there.

Karen Johnson: So, that's…

(David Behrens): No, that's a measure score, that's…

Michael Stoto: But I don't know what the answer is, but I think that's the kind of thing the measure would be.

(David Behrens): The measure will be performance -- yes.

Michael Stoto: Necessarily.

(David Behrens): And this is why we need to clarify.

Michael Stoto: Right.

Karen Johnson: Uh-hmm, uh-huh.

(David Behrens): Am I correct? I mean, the -- you take something like a PRO, like a questionnaire, it produces a score, that in itself is not the performance measure, the measure, that's an element that is used to come up with the measure score.

Karen Johnson: So, a measure score is the computed results of the measure, that other thing that I probably should have put in here is, when we talk about that usually data elements are most of the time patient level information, pieces from -- about a patient, right? The measure score is -- has been aggregated up, so you're looking at performance of a clinician, or a hospital, that sort of thing.

(David Behrens): I was going to say, that's the key thing why we're having two different uses of the word score. The score in your sense is for one patient, one number, the score for the provider, the entity is the aggregate, the average to somehow of multiple patients. And maybe that helps us sort out.

Male: Or maybe it's worth -- or maybe it's worth clarifying went in those cases where there are the number of survey items that are then aggregated into a single patient score which those scores are then aggregated into a provider level score, which is the data element in there.

Karen Johnson: Right. It's even trickier and I realize that I had -- I had not put the word composite on, if you're thinking about, and this is maybe even a little -- maybe what you're getting to, Andrew, maybe a little different. If you have a survey and five questions are all about this idea of respect, the individual five questions we consider the data element.

Those together may come up with some kind of a value like you were saying for a particular patient, a lot of people who do that kind of work, they would call those composites, just so you know, when NQF talks about a composite performance measure, we're talking about something different. We're not talking about those five elements that together represent a concept like respect.

And we have a formal -- we actually do have a formal definition of composite measure, because we pull them together in a group to help us come up with that definition, so.

Male: Just to clarify, Karen, the example that you gave, and I think that you mentioned so far, they can -- when the -- when you come up with the score from -- say, for example, from a instrument, and the instrument had five questions.

I think those individual questions -- they also be the data element, as well as the score it tells the items? We are talking about two levels of this?

(David Behrens): Yes, I've been thinking that we you know, we're probably going to need to have different levels of data element, so if there's not one level. So, yes, they are elements.

Male: Yep.

(David Behrens): And they are data. So, they're data elements. But then, even the score -- the score that that produces is also a still just a data element, right? And there are -- all that has to do with reliability at the data element level. So, we may need to layer the data element. And then we are…

Male: Yes, I mean, and -- that's what I'm trying to say. And also, limited this, right? How can I put or managed it based on an instrument, which is validated, how much do we need to get into the data element reliability?

If there are, like, instrument, like, scale or score reliability, right? If you want to get in the data element, it just opens up a whole can of worms, I mean, how much you want to dig in?

(David Behrens): We have Paul and then -- and then Susan, and then Mike.

(Paul Kurlansky): Just a quickly to clarify to (David's) comment, I think conceptually there's like a hierarchy of data elements, a way to think about it. And an individual answer or individual question to these data elements and the score that derives from that is a data element also, but it's a -- it's in a different hierarchy, if you will.

And there can be a hierarchy, it isn't necessarily, but there can be a hierarchy. And I think -- I don't know, from my point of view, there's a just working with databases, the concept of variables is a very helpful one.

Because what are the elements that go into figuring out where you, what your answer is, that's what you're looking at, but that element could be it could be an individual question or it could be the answer to a group of questions.

(David Behrens): If I -- and Susan, just one second, so -- and this is -- this could be weeds, so I apologize for that. But it may be useful, because we're talking now about

reliability, and just from my perspective, and we're using this example of a multi item scale, right? So, a lot of questions, five of them, ten of them, or whatever, and it produces a score.

If I was reviewing that for its reliability and validity, and I knew that it's the score, whether that's the PHQ-9 score, or this respect score, or whatever it is, that's what I want to know the reliability of, right? The individual items, I'm going to be looking at for validity, like, does it really measure -- are these questions really getting at what they say it's getting at?

It's a -- to me, that's not a reliability question, because the score that it rolls up to is what you -- what we're assessing for reliability. Is this -- do you agree with this? So, we have…

(Paul Kurlansky): So, you are assuming…

Male: Yes.

(Paul Kurlansky): …that those answers to those questions have been previous -- there are – not only valid for the measure score, but also are answered consistently, consistently by the same person.

(David Behrens): Well, no, no, no, not -- no, not assuming, no. No, no, no. You -- wait, hey, hey.

(Paul Kurlansky): (inaudible) survey my definition of this…

(David Behrens): Maybe, I mean, if I just clarify, because there is an alpha coefficient internal consistency, so I'm not -- I wouldn't be assuming that, I would -- that would be reliability.

But I'm not really looking at the item itself, I'm looking at the way the items together are consistently responded to. So, yes. But, yes, does that makes sense? Yes? OK. Susan.

(Susan): So, the only thing -- I was going to piggyback off of CQ's comment a little bit, I think if we have a measure developer that using a standard validated tool in its entirety, then our work's a little bit redundant, right?

On the data element, SF-36, am I going to reevaluate that? Probably not nor am I probably qualified to, to be honest with you, right?

But what we have to be careful of is -- and this is in the -- are they consistent in doing what they say they're going to do when they start taking a few questions out, right? That breaks that reliability and validity testing, so I think we all -- I think we all know that. So, where I get a little tied up here, and I think the idea Paul said, databases, so thinking about variables is really good.

And at first, I was thinking, well, if you validate the smallest unit of analysis, then if you combine them all together, you're OK. Maybe. All right. So, I think we do have to have that hierarchical sort of idea of are the basic elements valid, is the way you're combining them valid.

And then, is how you're rolling it up to the measured entity level valid, because you have these other, I mean, what gets lost in this, and I -- and I have to keep coming back to when I read some of these what's your sampling unit? What's your unit of analysis, right?

And is it the patient or the provider. And, so, yes, I think we -- I think that hierarchal idea is probably a place for it, but getting the variables defined well even on the form is a trick I think.

Michael Stoto: Two things. One is, I think that if we had had this conversation before that discussion we had a couple months back about what reliability meant and how to measure, we would have been -- that would have been a lot more satisfying, to me at least. I think the problem with it we just -- we had -- we all have different ideas about what these things meant?

The other thing I want to say, I just want to underscore something that Karen said, but I think to me the main distinction between data element and measure score is that the first one is that -- of that level of individual patients, and the second one is of providers. And that's the fundamental distinction, I think I'm getting some of that, is that true?

(Paul Kurlansky): I'm not…

Michael Stoto: Well, it's not the fundamental, these are important line of work.

(David Behrens): Well, I generally understood that to be soon, I was also thinking that in Susan's comment about the about SF-36, that it we -- it's not really a good use of our time to go all the way back to the 1970s and '80s, and John Ware, and all of that work that was done.

But it's certainly a fair question to say, is the SF-36 a reliable measure of quality when it's applied to say a group of orthopedic surgeons, and their outcomes, or a hospital, or something like that, because you can ask questions about reliability that are said in that context that would not necessarily have ever been thought of.

So, and maybe that's also sort of the weeds and hierarchy set of issues, but it to me, these things -- and I think that's sort of why I was not (inaudible) when you said about the measure score.

But to me the key thing is when you link that -- when you measure the performance of an entity, whether it's a physician group, a hospital, a nursing home, or whatever it is in the NQF context, you can use a measure or (inaudible) like SF-36, but you're using it in particular way you're providing a score, your imputing a quality leaning to it, which I understand leaks over into our validity discussion.

But anyway, I would want to -- when I first saw the term measure score, well, that's what I had in mind. It's the average SF-36 score of this entity for this group of patients, multiple patients in this context with this (inaudible).

Michael Stoto: But I'm not -- I don't think I heard -- I didn't see Karen nod when I -- when I said the distinction. I mean, so I want to -- OK.

Karen Johnson: No, I do agree with your distinction, the only kind of I'm tiptoeing around a little bit, because there may -- there probably are some measures out there where you're not looking at individual patient data, they might be -- and I can't think of any right offhand, but maybe not aggregated data or individual level data versus aggregate might be the better way to say it. It's usually patients.

Michael Stoto: OK. So, the measure of volume of surgery?

Karen Johnson: Yes.

Michael Stoto: I wouldn't have -- would be an example.

Karen Johnson: Yes, yes, yes.

Michael Stoto: But the volume would definitely be measure score that relates to the volume provided by certain hospitals or something like that?

Karen Johnson: Right.

Michael Stoto: Yes.

Karen Johnson: Yes.

(David Behrens): You still have a comment?

Michael Stoto: Well, I -- was it a higher level comment we quickly got into the weeds. So, but to go back up to that -- to go back up to the higher level comment, yes, it was I think the riddle of the inconsistency that we're trying to solve, if we have a very short list of a few things that -- at the root cause of the inconsistency of our ratings is this terminology.

And so, I want to we're talking about it right now. And I don't know if we're going to come up with the definitions for everything in every context right now.

But I did raise this to -- with NQF when I was going through one of the forms, the frustration with it, and the person from NQF at the tim, so I'm just given an example of it, I'm not picking on a person and said, "Oh, well, some of those things are defined in this this document."

But then when I read the definitions in the document, that definition did not seem to fit the context in which I was struggling with the term. So, it seems like we have to define, we have to define these things and it sounds like from

what we're saying right now, it's oftentimes probably pretty specific to that -- to that place on the form and it might change somewhere else on the form.

So, if we just really have to solve this riddle, and appreciate to that we're we're coming at it from very different contexts some of us are psychometricians that teach it, lot of us, some folks are clinicians, some of us are measure users, so there's -- this is just a -- I think a really big sticking point that we have to -- we have to figure out the -- and get the terms defined.

I'd like to see it really adjacent to each question when these terms are used, what is the definition of it? So, I think that would really go a long way.

(David Behrens): Well put, it's the glossary that we need and that's the toolkit, whitepaper, and whatever we're going to call it by the end of the day. But let's continue on, Karen, and I think that --

Karen Johnson: Keep going? Yes.

And I will say that you guys have brought up something this idea of the -- I think it's a parking lot issue probably, this idea of when you take these multi items still and create a score there that also is a data element.

And when we are thinking about what is the reliability or validity, I think we need more clarity about what we need to be looking at, because traditionally, we would have said, we want to make sure that question like (Ron) was saying that people answer it consistently over time at that item level, that's how we've thought about it, not so much, I think at that rolled up score level. So, that might be something that we need to add. So, let's make sure we get that on the parking lot.

(Paul Kurlansky): If I could comment just real quick on the data elements. Would medical records and the information in the medical record review, so -- if extraction, those are data elements also. And they represent…

Michael Stoto: Right.

(Paul Kurlansky): …quite different…

Michael Stoto: Yes.

(Paul Kurlansky): …kind of thing from like a survey or a diagnostic code OK.

Karen Johnson: And really you could see -- and it's kind of clear when you talk about our forms, our submission forms, the evaluation forms you guys are doing, even our definitions, we try to be generic enough that it's flexible to kind of fit whatever's coming in the door, because I guarantee you, we'll think of all the possible scenarios we can think of if we were trying to do a form for everything, and tomorrow something different will come in.

So, that's the other thing that I think we struggle with. So, we want to be as that sensitivity specificity business kind of comes to mind here with our definitions too.

Male: Yes, I guess with surgical training I tend to be very kind of great, but I think when you make the glossary, if you can put specific examples this is not meant to be inclusive, we're all inclusive, but certain examples would be particularly targeting things that have been identified as being potentially confusing.

Karen Johnson: And hopefully that will help with John's thing too, having examples. I'm going to talk about these assumptions really quickly, we talked about these I think in March in our call.

And I think in general, there was disagreement -- general agreement about these assumptions. Because our discussion didn't come back and kind of hit on those too much, but there's always going to be some error, and I really want to hit on this idea of reliability is not a static property of a measure.

And this is something that came up a lot in our emails back and forth, I think it's really important that we -- I know we all know it, but it's worth stating. That we know that reliability is not an attribute or at least, I think we know that it's not an attribute of a performance measure, right?

But that said, I know I try not to say that, but I know probably five times at least out of ten I do, because I get a little lazy or I get in a hurry, and that's

kind of the vernacular, so maybe one of the things that we can do is think about how we can consistently use our language so that we don't imply that we think that reliability is the property of a measure, it's a product -- it has to do with the context and the conditions in which it's being used. So if you can think about that for me, I think that would be useful.

Michael Stoto: Oh, it seems to that -- maybe a distinction between data element and measure score on this one here. So, for instance, on the measure score level, that's where the sample size matters, a lot. Where you could say that the data -- the data element level that that's not an issue at all, there's a difference.

Karen Johnson: Oh, Larry? Yes, uh-hmm.

(Larry): So, I have a comment about the statement that you made that reliability is not an attribute. And I think that when we're evaluating measures, so a particular -- say, we're looking at risk adjusted mortality after CABG surgery, OK?

So when -- if that measure is presented as a hospital based measure, where you have a pretty big sample size, OK? In that context, it may be a reliable measure, but if it was presented as a physician performance measure, because the sample size is a lot smaller, it may not be reliable.

But in my impression was always, when we're evaluating these measures, we're evaluating them as -- they're telling us what the unit of analysis is, OK? So, they're saying, this is going to be used for hospital performance measurements, this can be used for physicians.

So in that sense, it is an attribute, right? I mean, we should, I mean, if it turns out that the measure is not reliable for physician performance measurement, we should come back and say it's not. And it is an attribute of the way that measure is being presented to us. Does that make sense to you?

Karen Johnson: Yes and no, let me give our way of thinking about it and then we can open it up. What we would say in your scenario is, if you show reliability, let's say you looked at this data set for 2016, with these particular sets of providers and patients.

And it is reliable, there is not a guarantee that if you look at 2019 in a different data set that you would have the same reliability, the assumption is that if you can demonstrate that it seems to work well out in the wild and in this kind of scenario, it probably would work and be, and act kind of similarly next year with different patients, different providers that sort of thing, that's kind of the assumption. It's taking some stuff on faith, right?

(Larry): When I think of that, these measures that NQF is evaluating, I'm thinking that most of them are being used in a national level, and then people presenting us with national. With populate -- with relatively population based data sets. So, it -- so for example, if we're going to look at readmission measures, right?

Those are being used for all the if you're looking at AMI readmissions, that's based on a CMS data set, right? And it's not going to change from year to year, you're looking at CABG mortality, if you're looking at risk standardized mortality rates for congestive heart failure, it's the same data year after year.

If a developer is presenting us with a convenient sample of hospitals and patients, should we really be making decisions on whether or not that measure should be part of CMS' toolkit, to evaluate performance?

(David Behrens): Just a friendly amendment on (syntax) here, because I don't know sense that there's any profound disagreement here.

Karen Johnson: Uh-hmm.

(David Behrens): So, the phrase we have in front of us, not (inaudible) measure, seems to capture the middle ground that I think I'm hearing that what we're -- what I'm going to say is that the reliability is an the attribute of the measure across all contexts, OK? You're saying if it's you declare a context, you can say that reliability is an attribute of a measure in that context.

Male: So, I may -- even in the same context, right? If in that year the variation between facility decreases, right? So, nothing would change. So, even the same context they measure also depends on the variation among the facility?

Because when you calculate the reliability you rely on the between facility variation, so in the same measures, the same settings, same data, even participating facility, and they improve, and may get closer, now you won't be able to do a good job to separate them from each other.

Karen Johnson: Yes, I'm nodding because I agree, I think -- probably Larry agrees as well, I think also Larry, just FYI, you've had a chance I think in your work, you looked at measures that have come through that really do use the big data sets and the national data sets? That's not what we see across the board.

(Larry): I just want to be sure I got your point, I mean, are you -- are you giving a scenario in which basically there's change to a point where you reduce variability may no longer need the measure or is it -- is there a different point there?

Karen Johnson: I think also, Larry, just FYI, you've had a chance I think in your work you're supposed to add measures that have come through that really do use the big data sets and the national data sets. That's not what we see across the board.

Male: I just want to be sure I got your point. I mean are you give a scenario on which basically there's change to a point where you reduce variability and may no longer need the measure or is there a different point there?

Male: So it's possible but -- and that's what I think I tried to get to, you -- every time you report it, you maybe (inaudible) based on testing, I think you always quantify the uncertainties associated with your estimates. So in that, you always capture in live situations not how -- is the measure score functioning in the new setting.

But you can never get around that because my -- if you look at (AMI) mortality 10 years ago from now, you see the curve, I mean it's so much narrower. So if you want to calculate that based on 10 years ago on the same data set -- well, same data and model, yes, but it will be much higher than now.

Male: Just trying to decide -- see if you're saying reliability improves or performance improves and maybe it's performance, it's both, yes, yes.

Male: I think just a quick comment, I think -- I think -- if I could paraphrase I think what you're saying is that as the variability between providers decreases over time and it does, your signal to noise ratio goes away.

And so the reliability goes down. It has nothing to do with the measure performance. The measure is just as good as it was. It's just that providers have gotten a lot closer together.

(David Behrens): (Jennifer), is your current up, or you're just nodding -- OK. You're just -- OK, all right.

Male: In that setting your recalculating may be a little smaller, (inaudible)…

(David Behrens): Yes.

Male: …between variations is smaller.

(David Behrens): So (Gene) and then (Jack Needleman) on the phone has handout.

(Eugene Nuccio): In addition to this issue of improved performance that shrinks the distance between providers, if you are using a measure across settings, across different provider types, the prevalence of the condition, OK, makes a huge – has a huge impact on reliability, for example.

If the prevalence in one setting is 20 percent of a condition -- let's call it pressure ulcers. OK? And in the other setting, the prevalence is 0.6 percent, then measuring reliability in that second setting becomes incredibly difficult and incredibly misleading because of that shrinking prevalence. So reliability I think does have an importance if you're -- if you're taking that same measure and believe that one size fits all.

(David Behrens): (Jack Needleman)?

Jack Needleman: Yes, hi. I want to add another dimension to the conversation we're having about reliability, which is, it keys off of this concept of reliability is not an absolute. Colloquially, we've talked about reliability as is it reliable enough to be used for the purpose that it's being proposed for.

And yet we've had very conflicting guidance on whether or not to take use into account. So to be quite concrete, we've been looking at standard measures of reliability, signal to noise ratios, the inter-rater reliability or split sample correlations.

But many of the times, we're looking at measures which are being used for payment as has been noted earlier, where there's a very sharp edge cutoff. You will be penalized if you are in this percentile of the distribution, you will not be penalized if you're in the rest of the distribution.

And in those cases, the measures of reliability that we've been getting don't answer the question of how stable is the classification of somebody for penalty or not, what percentage do the folks move from the penalized group to the not penalized group if you change the data set or you split the data set.

So we've had very -- I think we've had very mixed signals from NQF about whether or not reliability needs to be evaluated in the context of use, because of the multiple uses to which measures are put.

I think we need to think about whether we need to be tailoring our assessment of reliability to the use and whether in – if something is being used for payment with sharp edges, we should have a different way of looking at the reliability than if it's being used for a general assessment or a relative ranking which is going to go into something like Hospital Compare with far fewer penalties involved.

Female: Thank you, (Jack). I think that's something that we want to put in our parking lot. It feels a little bit like our online discussion that we had during our break. It's something that NQF has tried to ask people not to think too much about.

And I think one of the things that I'm thinking about is we think about these big programs so you could talk about, OK, I know it's being used in XYZ program, but it may also be being used in a state or by an insurance plan or something like that.

So the same measures are often being used in different places, so I think that complicates it a little bit but let's put that on the parking lot. We won't lose that but it's actually a great segue I think to the next one. (NZQ), yes.

Male: I was the one who -- we have one question. You know, for measure development, you try to incentivize quality improvement, right? Now, if you choose a condition in our account (inaudible) most facilities, they are not doing as well as you hope for.

So they are performing at the same level, the level that's not ideal. So in that situation, your reliability is going to grow. We know everyone has room to improve. So just because the reliability is low because you can't separate from each other, but everyone is doing not as well, so how do you handle the situation?

Female: And that actually gets us way back to the beginning of our day when we said right now, NQF endorsement means suitability for accountability and (QI), your situation that you just talked about might be a time where we would say it's not really suitable for accountability, at least given the data that we've seen, certainly could still very well be used for (QI). So I think -- I don't know if you have anything you want to add to that.

(David Behrens): If I could just ask for clarification. I'm not sure I fully understand the point, but let me try a hypothetical. When you said that not all the performance is ideal or where you want or something, let's say that there's some measure hypothetically in which the way you want is a hundred percent.

There's something should be done all the time, where people should live all the time, whatever. And if I can just -- and you say you've got a set of entities, call them hospitals, everybody's at 70 percent. Is that a -- kind of fair statement of the situation you described? They're not different from each other but they're not ideal? OK.

This is something I was going to bring in anyway. Now is a good time. Some of the concepts of the reliability seem to have to do with -- they require variation among the entities being measured and I question them, at least from the background I come from.

Why could you not have a perfectly reliable measure that would give each of those hospitals a score of 70 and they would not be one jot different from each other? Why would such a measure not be reliable?

Male: I'm saying that if you go based on the formula, the reliability number will be low. So it's sort of not right. Sometimes it's not -- more than just a number. You know, if you really -- you want to calculate reliability, you plug in the formula, you will get a low score.

(David Behrens): Well, and I'm just -- this is -- I'm not sure this is big picture or weeds, but I know there are definitions of reliability that are like that and I don't accept those. I don't -- I don't think those are the -- they can be right but they're not complete.

Male: Right.

(David Behrens): You know, let's say I'm measuring the temperature of a hundred people who are perfectly normal temperature, they're all truly 98.6. I can ask questions about the reliability of my measurement but that doesn't require that they be different, these people and their temperatures.

I could have a reliable measure I think in my view of it where they all come out 98.6, there's no difference whatsoever. But, again, that's not to settle in one minute but I -- this seemed to be the time to introduce that--

Male: Well, can I just -- you mind if I comment? I mean I -- completely agree, (David), I mean but I think it gets complicated because there are so many types of reliability and ways to assess reliability but I think in your case, if you're a developer and you had a thermometer that was that good, whatever that is, then you should look at reproducibility.

And say that's the appropriate reliability statistic for my measure because it's completely reproducible and you'd have very good reliability. It may not be useful to -- for performance measure because it doesn't differentiate.

(David Behrens): Yes, exactly right.

Male: But it's very reliable.

(David Behrens): Right. And I'd say the measure is reliable, the ordering is not reliable. In my examples, the ordering would be essentially random, meaning once in a while you get a 98.7 and a 98.5 or a 69 and a 71. But let's say the measure is reliable, the ordering is unreliable.

Male: I mean I think…yes. So I'm just trying to draw the attention tonight, it's not -- it's not just number. And I -- based on certain formula where exactly I get you a reliability of one, believe it or not. But that number is highly unreliable. So I think we just sort of, we have to look closer at what information present in the form.

Male: My -- I think this has come up because I think it's really a fundamental point. And it gets at the issue that we just say, OK, if they do a signal to noise ratio and that's high, that's great.

But that would not -- but that could be a measure that if you could find a way to do kind of a test-retest thing, it would actually look pretty good. But that wouldn't meet that test.

On the other hand, you probably can construct a situation where you can find a group of clinics where there is a lot of variability and that makes -- it makes you look good even though it wouldn't look good compared to to really differentiate between the ones that we care about.

Male: It's like that story of looking for the keys where the lamppost is, even though you dropped them somewhere else.

Male: But I think it's important to, first of all, articulate these points and then secondly bear them -- bear them in mind that we're reviewing these things. That it's not that, OK, did a certain kind of test and then it got above a certain threshold and that's good. And otherwise, no good.

Female: I think what we're maybe talking about is the robustness of the reliability measure. Is it -- is there, you know -- I always go back to -- to me, it's more

intuitive to think about correlation and how if I have, if I'm over-powered, my correlation is always going to be statistically significant, right?

And it gets harder to be significant as I get -- as I get less and less power. And that's true with reliability, too, but I think we need to think of robust measures of reliability. Some measures of reliability are appropriate in a situation where we do have some between measure or between group variability. Others are better suited for other methodologies.

And that's I think we can't go with a one size fits all as I think we all know. But this is really a -- how robust is the reliability, is it transferable across settings and we -- it gets hard -- it's hard enough not to get reliability and validity tangled up, let alone reliability and how robust your measure of reliability is and appropriate your measure of reliability is, right?

We forget that all these statistical measures have these nasty assumptions under them and we don't like to think about that, right? And (inaudible) is our friend but it's not a -- it's not a skeleton key. So I think just statisticians get—

(David Behrens): That could be the title of an article. (Andrew) and then (Christie).

Andrew Lyzenga: I was just going to add that this gets to a point -- in the validity discussion, we do ask a question about whether the measure can identify meaningful differences among measured entities and we were actually going to ask later on whether that was distinguishable from this sort of the reliability question.

It kind of sounds like we're getting to something that is different, that maybe a measure is reliable getting a good signal to noise score but that it's not meaningfully differentiating among providers and that maybe gets -- maybe we should leave that in validity and it is a separate aspect. I just wanted to note that we'll talk about that a little bit later as well.

(David Behrens): (Christie)?

(Christie Teigland): Yes, I just wanted to share an example that I ran into with very argument that the developers gave, which was that just ignore the reliability score because all of the providers volunteered and they're all really good.

And so there's not -- you know, they're all the same. And so it doesn't mean anything that the reliability is only 0.2 or something like that. But you need to test it with the right data. If this is not a valid sample, then I reject.

(David Behrens): Some of those things we're talking are in these subsequent slides so we'll come back to them.

Female: And I was just thinking how smart you guys are because you're kind of anticipating the next slides. So you get that, right? So what we said is at NQF, so I'm kind of giving you the NQF version as is, knowing that we might expand or change somewhat.

We are thinking about reliability in terms of repeatability. In small words, I have some other -- in small thoughts, I have some other words so we can talk about whether those are synonymous or not. We generally use them synonymously.

We talk about repeatability and precision. And we also -- we very much agree that this idea of the concept of reliability can be applied to individual data elements as well as to the computed measure score, right? That's how we've looked at it.

NQF, up until now at least, has been interested in the repeatability of the data elements, right, and the precision of the measure score. And when we say precision, our definition right there, proportion of variance in scores due to systematic differences across the measured entities, the signal to noise idea. OK. So that's where we are right now.

What we've heard from you guys and I think -- I think this is part of what the discussion has been today, is we have this idea of repeatability, maybe that needs to be something that we need to think about for the measure score.

And maybe it's not quite the right word, repeatability, but getting there. And is there anything to be gained about trying to think about precision of the data elements?

This idea of repeatability of the measure score, the stability idea is new to us, right? Because up until now, we have said that reliability of the measure score is all about whether there -- the signal to noise is high enough that you could actually differentiate between providers. Right?

So that's what we have been interested in all along. And, again, it goes back to suitability for endorsement. We think that measures endorsed by NQF should be useful in accountability applications, if you can't different between providers, would they fit in a -- in an accountability application. So our questions to consider, I think I've…

(David Behrens): Can I -- can I just say? We have six questions to consider because there are two slides with three questions each. We have 45 minutes so about 7 -- I'm not -- I'm not saying we must spend 7 and only 7 minutes.

That's about the average just to keep that in mind. And I know -- (Paul) and (Mike), if you wanted to say something now or wait and see the questions because your comments may relate to the questions, you decide. If you want to say something now, go for it.

Male: Let me just say -- go ahead.

Male: Actually, I find it relates to the first question but I think it's very -- the word stability is very interesting. And I think it is distinct from repeatability because it introduces the element of time. And you may have a measure which is right now, very repeatable, very reproducible.

But there may be some reason to believe -- I don't have an example really, but there may be a reason to believe that it will not be stable over time.

You know, if you're measuring a performance of something which is radically changing right now because of technological advances or scientific knowledge or something like that, the viability of this measure as a stable -- may not be stable.

It may -- right now, it may be great. You know, but predictably three years from now and that's -- and if it's not going to be reliable three years from now

or it's predictably not going to be reliable three years or it's going to have to be retested every year, that is a major issue, which I think -- so stability is really a very interesting term. But I just wanted to make the distinction that it's adding a new element that reproducibility does not have.

Male: You want to go to the questions?

(David Behrens): It somehow went back to this slide.

Female: I -- oh, OK.

(David Behrens): There we go. (Mike), did you want to say something before? Go ahead.

Michael Stoto: Well, I mean it's actually – it does get at the first two questions because I want to talk about stability. And when I -- when I teach about this, I say that the concept is you can calculate in the month of April what proportion of patients with a heart attack got an aspirin.

What we really want to know is if you could re-run the month of April in this hospital with the same staff and everything else, same procedures and so on, how close would that number be? Again, I think that's really the concept we want to get at. It's not always easy to get there.

You know, we -- when we -- we have the binomial distribution, it tells us something about that based on the sample size. I think that's the concept we want to get at. But the other thing I want to say is, (Andrew) made an important point I want to come back to.

You talked about meaningful differences. And I think that gets back to the conversation that we were having here, we were talking about how much variability is there relative to some distribution of measured things but we have to bring in this concept of meaningful.

Female: And we can parking lot this if we want but I have a pretty -- I think a significant issue with equating precision with reliability, because precision is a concept -- is a sampling concept and it's about variability of the measure. It's not about reliability at all.

Reliability is exactly what you're saying, (Mike). I do -- if I could recreate April or whatever month it was, I want to recreate April, right? Precision is if I sample a hundred patients to see if they have aspirin or not. Is my margin of error narrow enough that it's useful, right? So I don't -- I don't know -- does anybody else have…

Male: (Off-mike)

Female: OK.

Male: I mean, I think it's -- precision is (inaudible) standard deviation. It's very critical concept. So we cannot say definitely but I guess -- I mean I agree what you were getting at.

Female: And just let me add in and then maybe (inaudible) is going to help me out here but I think we think about it because if the standard errors are small, we probably will have a better chance of differentiating between providers. That's the link that I'm seeing or you're kind of -- so maybe I'm not quite right but --

Male: I think the term people interpret differently -- I think particularly for statisticians, they won't like it when you say precision and you call it reliability. Your description in the slide actually is appropriate.

If you go back to -- when you say the precision, you had a description I think that sort of more captured the signal to noise aspect on there. So that's why we struggle. Sometimes people interpret it based on their background so it's…

(David Behrens): So yes, I just would like to make a general comment, perhaps restating the obvious or the things that we've agreed to in the past but seeing them here -- so much of the challenge that we're having is related to terminology and what words mean to different fields.

Precision, to me, is reciprocal of error. It refers to the amount of information you're getting. So it is sort of a signal to noise thing but I think -- work it out as a reciprocal of error, but I'm sure that's not what it means to somebody else.

And so maybe we need to talk about what we really need for the measure and put it in lay terms and then come up with a word for it that we can -- that we can live with as opposed to starting from the words and trying to get everyone to agree.

Similarly and I'll just say from my background, repeatability, reproducibility, stability, what's the other one -- I don't know, not consistency. But repeatability, reproducibility, and stability are all the same thing. They're synonymous, but I can appreciate -- I know, (Paul), because what you saw was time.

And I'll tell you, I think I know why. Because all of those are measured with a test -- in my world, with test-retest. You have to have two points of assessment. But what I see happening here is that people kind of do a shortcut and they say we're going to cut this sample in half.

And if there is -- they're the same, that's probably a sign that it's repeatable or stable or reproducible but not necessarily -- but that's what's done. And so one could develop an idea that I'm going to call that reproducibility but that's different than stability, which I think may be where--

So, again, I think it's just an example of how -- maybe what we need to say is let's not work with the words because the words keep getting us tied up, and focus on the, terms rather and focus on the language that that relates to the measures that we're trying to get our heads around. Makes sense, yes?

(David Behrens): I think -- you want to follow up, I think you and then Larry?

Male: So I think we're going to run into this more often in the future. So when we talk about data element, I think there are two big categories. One is sort of more objective, honest measurement, right? (inaudible) EHR-based measure and that one is more subjective. Even like say (inaudible) some data element whether you have this condition.

So it's a different way that we'll get the same thing, right, that's kind of subjective. But some others, so I call evaluative. So that's gold standard.

You want me to be as close to that (inaudible) sounding more subjective. So I just -- I can see that when we have more measures based on EHR, we're going to have that issue.

Male: I love this discussion. So I think it's really easy to get stuck when you're using words. I completely agree with (David), you know. And I really honestly start to have problems wrapping my head around all these different terms that we're using.

I think it's a lot simpler if instead of using -- we sort of say, OK, we're going to look at reliability. So what are some of the different ways do we have of quantifying reliability? So instead of starting with definition, let's just look at some of the statistical approaches that we have.

And when you do that, it becomes extraordinarily concrete because you don't have a million different ways of doing it. And they are very concrete, right? I mean if you're looking at measure reliability, OK, you have signal to noise ratio. And then you have your inter-class correlation coefficient using the test-retest approach. OK.

That's really -- that's it. I mean those -- and if we're going to -- you know, this afternoon when we're going to be talking about looking at scientific validity of measures, OK, so we have measures of discrimination and calibration and you can talk about those.

You have the C statistic, you have the Brier score, you have the Hosmer Lemeshow statistic. You have the calibrations -- when we start to confine our discussion to the actual how we're going to operationalize evaluating reliability, how we're going to operationalize evaluating model performance, it becomes, I think, much more doable.

Very concrete, OK, versus trying to look at the words and trying to say what's -- that's -- let's just look at the tools that we have to use to evaluate these different concepts.

And then let's make some decisions as a group, which tools we want to use and what criteria for those tools. You know, if you have a C statistic of 0.6, is

that good? I mean to some extent, make it a little bit concrete because once we can come up with that consensus, it's going to make it a lot easier for us to evaluate these measures.

And as importantly, we can give the measure developers feedback. This is what we want. We want you guys to use these two tests and these are the numbers sort of that we're looking -- you know, some of the ranges, realizing that it's kind of hard to come up with cutoffs.

But just to give people some sort of an idea. So I would move that we make this more concrete right now. That we talk about which criteria we're going to use to evaluate reliability how are we going to do it as opposed to sort of spending too much time with the terms themselves.

Male: OK.

Male: OK. Friendly amendment to this perhaps? There might be more than two. I'm thinking there might be seven or eight that -- and the selection among them might be appropriate to different circumstances.

I can think, for example, in a situation we had an opportunity to measure energies twice near to each when you knew that their performance was stable underneath, a test-retest would be very compelling to me.

I can think of other situations where I want to see an interclass correlation coefficient because in that context that would be the best way to assess the concept. So I -- and rather than debating is one or other somehow the more faithful capture of reliability so here are a menu of choices basically that live under this broad concept.

And we might in this white paper think say in general here are the circumstances under which test-retest would make sense. Here are the situations under which (inaudible) might make sense boom boom boom, do what makes sense and that's what we're going to be looking for when you come back in. I mean is that in the spirit of your comment?

Male: Yes.

Male: OK.

(David Behrens): (Gene) and then (Lacy) and then (Mike).

(Eugene Nuccio): Probably following on with (Larry's) comment, but I think even more fundamental question which I think we keep forgetting is are we talking about reliability of the data element or are we talking about reliability of the measure which is the compilation, aggregation, averaging whatever of all those data elements.

Until we begin separating and thinking about repeatability, reproducibility, and stability of the data element, OK, and talking about the data element what kinds of measures are we looking for at the data element level, maybe we're talking about a (CAPA) if we have a survey. But if you're taking an extraction out of a medical record, you're not talking about a (CAPA).

So, the tool that we're using has to be fitted to whether you're talking about data element or measure, and even within measure, are we talking about a regular outcome measure or are we talking about a composite measure.

And is there a differentiation are we going to make between a new measure coming to us or a measure that's already seen usage for five years kind of thing?

So, I'd request that that we keep thinking about data element and measure when we're talking about reliability and validity because there are different questions that need to be answered – fundamentally different questions that need to be answered about the data elements reliability versus the measure reliability.

Female: Hi, (Lacy Sabian). So, I agree. I really like this discussion. I think it's good and we're getting somewhere.

I just want to add on to the piece of this consideration for getting more concrete, tying it all back up to our overall charge and that these measures can be used for quality improvement and accountability.

So, that's going to dictate, right, some of the methods that we can use and what our expectations are for what those analyses look like based on how we're going to actually use the measures. I know if we're going to start tying something to payment, we must consider what we're expecting out of these measures.

And to Eugene's point, I think so far in our reviews, we have looked at measures differently if they're up for re-endorsement or maintenance versus if they're completely brand new, as well as what is happening in the field in that particular measure concept area.

Because some concepts have a lot of saturation. So, the bar might be higher with what's expected of the measures whereas other conceptual areas aren't quite as saturated, so the bar might be lower, just additional considerations. I'm all for the concrete.

(David Behrens): And then Jack on the phone.

Jack Needleman: No. I'm going to respectfully disagree with Larry. And I think that if we start by saying "Here are the ways we have of assessing reliability and so on", we will some of these subtle points that you bring up and so on and the idea about test/retest versus – what's that – yes, split half and so on, they really get a different concepta. I think we really have to be clear about what are the concepts that we want.

So, David, when you suggested that friendly amendment I think it actually was pretty fundamentally different, that we think about how different measures are used in different settings and the way to resolve that is by going back to the primary concept.

(David Behrens): Yes. And again, I always try to be friendly, but I think we don't have massive disagreements here because I think where we would end up would be a situation where we as a group would say there are multiple choices that one can make as a developer to assess this concept of reliability that the different choices are appropriately selected for different circumstances that we can identify.

And that I sort of then end up where we're willing to accept information on one or two or three of those. They appropriately selected ones and I'm sorry. I'm not closing very well, but…

(Eugene Nuccio): But I do think we have to say – we have to recognize that signal-to-noise ratio isn't always a thing to look at and really that only gets at a certain aspect of reliability and when we need to be clear about which of these things are appropriate and when to get there is by thinking about first principles.

(David Behrens): So, I think it was a friendly amendment. It did come across very friendly and it did sound different though.

And I tend to agree actually with apologies to Larry that starting at the statistics could get us into real trouble because I would envision then seeing a submission that says, "OK. Here's your ICC" and it would be the wrong statistic to use in that setting. So, we should definitely describe the statistic that is relevant to a setting, but we need to put that context in there.

And I'm actually thinking and Jack will (come) in just a minute. I'm actually thinking maybe the way the review is set up now, it kind of goes from the specific to the general, it builds. Like, is it reliable? Is it valid?

Is it a good measure? What about starting from what's the measure and what's the evidence that the measure is good? And then, deconstructing the data – have the developer sort of deconstruct from the measure down to the data elements with appropriate statistics sort of defending that deconstruction.

This is maybe a way to think about it. You start from the measure, the overall measure, the general and then work your way down a pathway to the data elements. We're looking at statistics as they support the position on this is a good measure. Anyway, Jack, you're next.

Jack Needleman: Thank you, great discussion. I agree that one ought to be thinking about what statistic makes sense for answering the question about reliability and the context of use. But one of the things we're seeing in the presentations we're getting, people are picking statistics, an ICC or signal-to-noise ratio.

And they're not only picking their statistics, but they're going back to the textbooks on whatever number they've come up with and Larry was quite right about picking numbers I think, because we're seeing them.

Actually, people are presenting them to us. They're going to the textbooks to say is "The number we're getting from this measure, how do we characterize it? Is it excellent? Is it good? Is it highly…"

My sense looking at some of the measures and particularly some of the ICC measures, the split sample measures is that if again, I'm going back to my sharp edged samples, there's enough movement across quintiles or quartiles or deciles that what looks like a good measure by the standard textbook number is not good enough for use, for the intended use.

So, as we begin moving down the road here, I'd like to see us actually looking more closely at what – and this again may be a parking lot issue for the moment, but I'd like to see us think much harder about what actually if we're going to see the statistics with numbers, what numbers are good enough for the use that we're actually seeing and the measure being put to.

And I suspect that the textbook numbers are too low at this point. So, that's why I think it's an important issue for us to think about.

Karen Johnson: Thank you, Jack. So, these questions, we may not get to all of them. We've kind of danced around I think some of these. I think the one that I really want to understand is the middle one. So, let's start with that one, this idea of stability.

And I actually do want to go back to Gene's point and I want to remind everybody that for most of our measure types that come in the door, not all but most, our NQF requirements say that you can show us data elements reliability or score level reliability, right?

So, again, some of them we say we have to see both. But most, we say it could be either or, right, and it's dealer's choice. People can bring in what they want and it often depends on what kind of data they have available.

When we think about data element reliability, again, trying not to get too much in the weeds, we've mostly thought about for most kinds of measures, we thought about the abstraction piece. So, can different people abstract and would pretty much come up with the same kind of thing, right?

So, that's generally what we've thought of. It's a little different when you think about reliability as the items in an instrument. We're just going to put that off and agree that that's a different animal.

So, that's what we've looked at for data element reliability. Again for measure score reliability, we've been interested in can you differentiate between providers.

That's typically some sort of a signal-to-noise and gets a little bit to this idea of the bottom bullet where I kind of used this signal-to-noise as a short-hard way to think about it. We might need to work on – that might not be quite right especially the statisticians may really be feeling uncomfortable there. I'm not sure.

But going back to this idea of stability and I think I want to limit this idea of stability not so much over time, or at least not over much time, because we have said and developers have brought in, they say, "OK. Here's the performance score this year and here it was last year and the year before and the year before and it's pretty stable."

And we've always said "I'm not interested in that because quite frankly, we don't want that number to be stable. We want it to get better, right" and we hope that it does. Stability within the same population, the same timeframe roughly is I think what we're trying to get to here.

So, the question for you is this idea of stability, it does feel new to NQF. It's different than the differentiating between providers, right? So, our question is, do you feel like the stability is just as important as being able to differentiate between providers?

And the reason I'm asking you is if you think that's the case, then, we would probably start saying, if you're going to tell us about measure score, you have to tell us about both of them, right? Is that clear, my ask of you?

(David Behrens): I think so. Could I just say, at least (Matt and then David), but just to my suggestion earlier about dissecting the word, so in this context, are we defining stability as the extent to which the measure – whatever, data element, does not change during a period of time where it shouldn't change, forgetting about the one-year example.

I understand that, that you should be improving in a year. That should change. But if you take a period of time, it might be a week or a month, it shouldn't change and I think this is what some people estimate with split half. They shouldn't change.

Therefore, it's stable in a situation where you don't expect it to change. And the reason that's important is because if it does change, then, it's error and you have the signal-to-noise problem. So, ultimately, it contributes to the noise and the ability to detect the signal within it.

So, I think that's the framework we're talking about, right, and you're asking is that important and is it as important as differentiating. And I think it's (Matt) and then David.

(Matt Austin): Yes. And, Karen, maybe you could provide me with a little more background on why NQF has landed on allowing developers to either demonstrate data element reliability or measure score reliability, because when I sit here, I get very uncomfortable with just doing data element reliability because I'm like I'm not sure that the final product is actually reliable. So, that's where I get sort of a little uncomfortable.

Karen Johnson: Great question. I wasn't around when those decisions were originally made. But my guess is that a lot of it has had two pieces. One is burden. We don't want to make things so hard that nobody comes to NQF for endorsement. So, that's kind of one piece.

Another piece is depending on the data that are available for testing, if you're only able to do some testing in three or four hospitals with a small sample size of patients, doing a signal-to-noise may not be possible or it may not – it just might not quite work for you and kind of the flip if you're basing your measures on claims data, being able to go back and do medical record kind of checking is probably not necessarily possible.

So, they're both very important and we would love to have both but I think there's some possible realities that might make it hard to do both.

I will tell you that three or four years ago when we built our algorithms, we actually started giving the score level testing a higher rating potentially than data element, kind of signaling that idea that we really want to make sure that we can differentiate between providers. Before that, what we were giving is moderate if you had one kind and high if you had both. So, that's kind of morphed over time.

Male: OK. I would second that concern. I'm actually surprised that the either or option was available and I share the reservation if you only came with data element, you're not there yet. You're not ready for endorsement.

On the issue of stability, it's funny that the word to me up until you clarified inevitably implied passage of time. Now, I think Paul made a similar comment.

So, when I was looking at this until you made the clarification, I'm going to say, well, I don't know if I'm a big fan of the concept of stability if it has this over time concept for reasons people have already stated.

You should see change. You should see within our organization, you could see movement as legitimate. The fact that there's movement doesn't mean the measure is not reliable. The fact that the ordering among organizations changes over time, that doesn't mean it's unreliable.

OK. So, now, you've taken passage of time out of the picture. Now, I'm not sure what's left for this word to have some unique meaning. So, I'm now

asking a question. Once you've taken that passage of time out, what's the difference in meaning between stability and signal-to-noise? Is there any?

Karen Johnson: I think it is and the – I think it is, but let's delve into it and see. I think (Christie) in the call was very concerned about kind bebopping back and forth. So, a plan that kind of comes in here when you do it. If you do it next month, it might be over here and again…

Male: But that's passage of time.

Female: Yes. But some things probably won't change much. Now, and the split half kind of gets to that because it's the patient group at the same time – sorry, it's a random sample. So, you can get around that a little bit. But I think the idea is that if there hasn't been change in behavior of providers, then, you would expect their results to look fairly similar.

I mean, you're going to have different sets of patients. But I think that's the idea. I have a couple of really good quotes that could read off from various email exchanges I think from Jack, and Larry, and maybe Mike that probably state it even much better.

Male: OK. Well, just a quick drop on that that point, the (inaudible) would have related to the main bullet within the first sub-bullet is that in either of those contexts, there has to be some accompanying idea of what the true reality is if we could know it.

So, for example, some of what you just said implied that the thing is actually stable, if somehow we had God's wisdom, we could see it stable. Then, you can say, is the number we're looking at stable. OK.

And I'm OK with that. But the same thing about distinguished differences, that's kind of the point I made a few minutes ago. I'm willing to consider a measure reliably even if it doesn't distinguish any differences if there are no differences.

So, in both cases, you have to say, what's the underlying reality? And then, if you can declare that, or have some independent knowledge of it, then you can

go look at some statistical test of reliability and set it against what you're assuming.

If you're going to subject a measure to it, a test of being able to distinguish differences, you first have to establish there are real differences to detect. If there are none, then, that's not a good test.

And if you're talking about stability over even a little period of time, you have to somehow declare or assume that there is actually stable performance in reality and then you can ask is the number stable. But in both cases, you have to somehow know what's true and I, you do or you assume it.

(David Behrens): Right. We do a lot of assuming. That point was made earlier by Susan with that great article title about the central limit theorem is our friend, but not a skeleton key.

I mean, to me, I'm sorry. We're going to go around and say one thing. But I have (Christie), Larry, Mike, and then, Susan and saying we were up and then down, but you're back up and Jean is back up.

So, we have a lot of people things to say, but to me, I would not want to take away the ability of a developer to show that a measure is stable over a short period of time and therefore helps in confirming, in establishing its reliability in that context.

I don't know why we would take that away. It seems to be a reasonable option, sort of a standard option for showing that the noise is managed by this particular way of gathering. I mean, if somebody is treating my depression and I take a depression scale and then I take it again tomorrow and my answers are really different, I haven't really gotten better in one day.

That's a problem and to use that as an outcome measure, but if my answers are pretty much the same, you've got stability. And then, later, you want me to be better, then, I'm better later, then you've got it.

So, to me, it's a good measure of the noise or a counter-reactant of the noise. So, I don't know if we want to take it away. But I read that and when I first

saw it, I said, well, you're asking if one of the things I use to establish something else is as important as it and I think of it as a component of it or at least a potential component. Do you see what I mean, right?

And I wouldn't pit stability against stability to distinguish differences. We use it to give us comfort when we establish differences. Anyway, (Christie) and then I think the order would be Larry, Mike, Susan, (Zhenglie) and (Eugene).

(Christie Teigland): Yes. So, I am thinking of stability in terms of time and I'm thinking about it more in terms of the whole way these measures are used in practice, especially for public reporting and that is if a health plan is a A, five-star this year, people are making decisions on that. They're getting paid based on that.

People are choosing that plan. And then, all of a sudden next year, they're a D. They're a two-star plan, a three-star plan. That's not good, right? There's something wrong with the measure. You wouldn't expect them to be, change that much over time.

So, we want some kind of stability. But then, you start to think about some of these measures where there's very little variation in the scores, right? So, it could be a two-point difference in a score where you're a five versus a three-star plan.

And I'm particularly thinking about like the satisfaction measures or like the caps measures where there's very little variation in responses. It's very easy to – because they take a small sample of patients they interview and to get a really pretty different score, at least four or five points from one quarter to the next, one year to the next.

And so, that's what happens. A plan can be a really wonderful five-star plan and then, the next time they get do a satisfaction survey of their members, they're a two-star plan. So, that signal-to-noise, the underlying reliability whether it's signal-to-noise or something else is really important in that case, and you're not going to see very much stability then over time.
And so, the people who are trying to use that measure for picking a plan or for paying that plan or for whatever are not going to have the base validity solidity that that is any use at all, that there's any meaning in that score at all.

And we want these to be meaningful and useful. So, let's not forget how these are used in practice.

(Larry): We're having a great discussion. So, I think that at first glance, the idea of stability, the reason that that's somewhat attractive is that you want to think about the goals of performance measurement. One of them is transparency and one of them really is a very patient-centered thing.

So, if you're a patient and you want to select which hospital you want to go to, you would like to know that the hospital performance in 2017 somehow carries over to some extent in 2018, OK?

And assuming that there hasn't been this massive change, usually, things are somewhat consistent over a relatively short time interval in terms of hospital performance or physician performance as long as you have a big enough sample size to be able to look at this.

And so, that's the reason maybe to look at stability. Although, I think stability should be a second order evaluation. I don't think that that's as important necessarily as some of the other things that we've been looking at.

But to again go back and being very concrete at the risk of – the way I think about this whether you're measuring stability or – I mean, there are two different ways of measuring reliability, right? One of them is the signal-to-noise ratio, OK? And the other one is to basically take your – say you're doing a split sample.

You're splitting your dataset from the same year, 2017, CABG patients, and you're randomly splitting it into two halves and then you're evaluating provider performance in one half and you're evaluating provider performance in the other half, and then you're using the interclass correlation coefficient as a measure of agreement.

And the idea is that, look, you randomly split the data. It doesn't matter whether your performance is evaluated with one half the data or the other half of the data.

It should be more or less the same, right? And so, that's kind of an important concept, right? I mean, if you score differently depending on what half the data then the measure doesn't work. So, that's a really key concept.

Now, stability is kind of the same thing almost the way we're seeing it. Basically, instead of splitting the data from 2017 into two halves, what you're doing is you're taking the data in 2017 and you're evaluating all the providers in 2017 and then you're doing the same thing in 2018, then you're comparing how they do.

OK. So, it's almost the same concept, right, except that what we're doing in one sense is we're just looking at within one year versus on the other one, we're using the same statistical testing and we're comparing performance from one year to the other.

And so, they're almost evaluating the same thing. And I don't really know that you need to evaluate stability per se. Because the way we're currently doing it with the split half technique is very similar and almost gets to the same point, right?

And again, if you wanted to do stability as sort of extra credit to show that, look, this is a good measure because the consumers, the folks out there who actually are using these measures, not just CMS, but the people who are actually using them and are using the report cards, we're showing you that you can use our measure for public reporting and it means something because you can pick your hospital depending on how well they did a year ago or two years ago in terms of making your decisions where you want to go.

(Eugene Nuccio): So, I'm going to agree with Larry. And I think that to me, what we care about is that if you could rerun the month of April, you wouldn't get a radically different number and that if you actually made an improvement or you're comparing A to B and one is better than the other, you know that as well.

And I think that there's a lot – a number of different ways we could assess that. One is split sample, things like that.

Another one is looking at it in April and May presuming that nothing radically is – nothing is going to change radically different in May. And then, I also think that signal-to-noise ration is another way of getting at that in some circumstances. But I think they're really getting at the same concept.

Some of those things may be more or less difficult to actually calculate but when we're talking about the reliability of the score measure, there could be different ways of getting there as long as we have that same concept in mind.

Female: Yes. I'll be really quick. Sorry. So, I think we're – I do kind of think we're talking about the same thing. So, if we think about the variability in a measure, we've got what we call the signal and the noise, right, the error and then the measurement of the difference between the providers or whatever entities.

When we think about reproducibility, stability, any of those terms in the first bullet, what we're thinking about is if I change the context or I change the timeframe, is the amount of random noise, right – is the amount of random error or noise stable?

The signal might change as people get better. And, Karen, that's what you noticed. I don't want it to be 2.5 all the time. Well, it will be 2.5 all the time if all we're – we want the random part of the error or the random part of the noise to be 2.5 all the time and if there are performance changes, we want to be able to measure that.

And so, I think believe it or not, let me say I think we are kind of talking about the same things and these may be approaches to measure reliability. You can demonstrate repeatability.

You can demonstrate stability in terms of random error. So, I think they are all necessary and probably sufficient, but I don't think you need to measure all four. I think they're just maybe bullets of ways to measure reliability.

Male: I just want to point out, I think you and I are talking about the same thing. Larry and Mike are talking – so, you and I are talking about the denominator.

Female: Yes.

Male: Larry and Mike are talking about the numerator over the denominator, the measure itself. And so, I think we need – they're both – I think the denominator is just part of what helps you have reassurance that the numerator or denominator is not going to create some of the false negative that (Christie) was concerned about. But I think they both are needed.

Male: Thank you.

Male: Yes. So, I think what Karen is trying to say when you talk about stability, I think in some way you are referring to the sort of parallel form test, test, retest inside psychometric has some specific meanings and that's why it's causing some confusion.

You are saying that you had two equivalent tests. You test them. You get the same, similar result, they will never be the same and that's why they will be different. So, that's why I think it becomes relevant. You need to have both numerator and denominator.

It doesn't matter what kind of activity we're talking about, and if you really pare them down, the eqution is always some kind of between (inaudible) and divided by between (across errors). So, in a way, we are always trying to get the same thing.

(David Behrens): The co-chair actually had, felt he had to put a card up. Can we just let Eugene go and then -- because you kind of have the right to just speak whenever you wish to.

Male: You're right, the last word. I'm putting my developer hat on right now. When I first saw the term stability, I was thinking about over time and the way a developer would use – would deal with that is you take a look at the ranks of an agency at time one, but the ranks of those same set of agencies at time two to and you split them into deciles files so to get rid of that little noise thing so that I should only be seeing signal. And then I'd do the nice little correlation kind of thing.

However, I would suggest that if we're really thinking about stability also, I mean that's OK, but a different way of thinking about stability as stability now, OK, that in the world that I live in, a way of doing stability would be to stratify my data across, perhaps, states.

So if I have the same shape bell curve of the distribution of that metric, that measure four, improvement in walking and I show that to be true for agencies in Minnesota and agencies in Florida and agencies in Arizona and agencies in Washington, then I would suggest that the measure seems to be working in a rather stable way across multiple groups and that thing.

So I was suggesting that we could think of not just split half, OK, but also stratification as a way of demonstrating temporally similar stability.

Female: All right, so my question for you..

(David Behrens): You have three minutes.

Female: Yes, three minutes, so what I think I'm hearing is at least (Dave and I think Susan) and maybe a couple of others feel like that this idea of we've tagged in stability.

That might not be the right word, but this idea of if you look at it at the first week of May and the second week of May, it's a little different provider patients that we should get roughly the same performance score for your measure, you kind of feel like you're getting to the same thing if you're doing signal to noise.

So where I'm confused and where I need somebody to help me understand is the signal to noise to me tells me that I have a greater likelihood of being able to differentiate between providers.

If you tell me that in the first week of May, I get roughly the same score as I would get in the second week of May, I don't understand how that tells me that I can differentiate between providers. So if you could answer that, then I would understand why they feel like kind of the same -- you're getting at the same thing.

(David Behrens): I could (inaudible) to address the issue of where you framed it I think you have to have this assumption of are there or are there not meaningful differences among providers and it's hard to think about it without.

The signal to noise thing to me is it seems like a broader conceptual concept as opposed to a specific statistical test or I'm not sure how I would describe the conceptual territory and (inaudible) just in general.

And I guess the way my training is to say yes, a particular measure has been able to show a high ratio of signal to noise. With that information then, you should be able to detect differences among providers if there are in fact differences among providers.

Now, stability, maybe one way to think about it is it's a way of expressing the broader concept of signal providers as opposed to doing a different thing and I'm looking around the table to say wait a minute, in the land I come from, there are two different tests that address those different issues, but then it goes back to my question when we started this, what is the difference between stability and signal to noise if over time it's taken out of the picture?

Male: So I think they are two completely different concepts which is I think your point. Yes and I completely agree with what you're saying signal to noise is -- I mean is there -- if you compare the variability between providers versus the variability within a provider, hopefully the variability between providers is going to be a lot bigger, so that you can tell the difference between different providers. That doesn't get at all to stability.

Stability is, as we talked about it, is basically you calculate the performance in year 2017 and you compare the provider performance in the year 2018 and it's very similar. I mean that's kind of the stability piece.

(David Behrens): And I'm with you that they're different if stability has a tie to passage of time.

Male: Yes.

Male: Is it fair to say as stability is getting within facility variation? I know you're trying to get at that, but it doesn't speak to the between in facilities?

Male: Is this weeds now?

Female: My definition of weeds is to talk about the formula for ICC or something like that. So we've managed to stay out of the weeds so far. I think maybe that is -- I think that's kind of what we're thinking and that's a good thing.

So it takes us back to the question, so I feel like I still don't know because I feel like some people think it's kind of the same concept.

Larry and I think they are kind of two different concepts, so we have to figure that out and if it's two different concepts, the question comes back to you, we definitely have to have signal to noise and we'd like to have stability or whatever we call that or we absolutely have to have both.

Male: I don't think there's so much disagreement as there is -- you know, looking at different parts of the elephant. So what I'm -- let me give you a very concrete maybe silly example.

Let's say you want to follow four year olds to when they're six, right, and now you know they're going to get taller, right? The six year olds are going to be taller, so you know that information to your point, (David).

And you have two ways of measuring height. One is a ruler, right, you know how good that is, it's very reliable, very stable and then another is this like spring thing that you have to like put against the kid and it keeps springing back and forth and you have to make your best guess.

That second measure is not very reliable, not very stable because your estimates might be -- there is a lot of noise. So you might not pick up that height change from age four to age six with that measure.

So, I've been talking about that measure as if comparing a ruler to the spring and I think that's what Susan was getting at. Other people are just talking

about well, we need anyway a measure of height and we need to see if it's stable.

And by that definition what Larry said, the spring one might look more stable but it's garbage, right, because it's looking stable but they've actually changed. So we're talking about different parts of the elephant, different layers of the issue, all we've been talking about I think is how reliable.

That to me is what reliability is. The ruler is reliable. The spring thing isn't reliable and as a result, you wouldn't want to use that spring in a measure of height over time because you wouldn't pick up the real change.

And (David) has been talking about how -- if you don't really know whether there is or there isn't change, well, you're working in an assumptive world which is also another problem. So it doesn't look like it helped looking by your expression.

Female: It didn't quite help because I think what I'm really interested in knowing not so much is did the kid get taller from four to six.

I'm more interested in knowing if I measured Johnny versus Suzy, can I tell a difference. I think that's -- so that's why I am still having trouble and if you're kind of tired talking about it, just let me know.

Male: So this is -- I'm going to stick with my example of getting aspirin when you have a heart attack. In DC, if you have a heart attack, the ambulance would take you to Washington Hospital Center rather than the Georgetown because that's where the treatment – they're set up to do that treatment.

And so they have a lot of people who come with a heart attack to the Washington Hospital Center and you could know -- so there's really very little difference from month to month because the sample sizes is relatively large, OK, unless they actually change their procedures or do something like that.

At Georgetown, there could be big jumps from month to month in terms of the proportion getting their, what do you call it, aspirin just because the sample size is very small.

So you really can't use that -- you would say there would be a lack of reliability in measuring that. So if you wanted to be able to judge the quality of the hospitals, you can do a better job at places like Washington Hospital Center than you can at Geogetown.

Male: So I think that some of the confusion I have here is when we talk about stability, it's sort of like we're talking about it over a different time intervals versus the when people talk about the split half tech approach where we're using within the same time interval but you're randomly splitting the data, but the same concept in terms of -- at least it's the same statistical test, OK?

So there's that one measure of reliability typically and without getting into the weeds, the cross correlation conversation.

And then there is -- and well, I shouldn't even use that term, but anyway. So there's that approach and then there is the signal and noise ratio which is very different and it also looks at reliability.

There are two different ways of getting maybe as you were saying earlier, maybe it's the elephant, we're looking at it in different areas. In the same way that if we were looking at model calibration, there're lots of different ways of assessing that.

And I think at the end of the day, maybe it ends up being an empirical question how closely related these two different approaches to measuring reliability are or are not. But my sense is they get at slightly different things and I think that going back to first principles, I think that we do need to look at measure reliability I think for every measure, OK?

I don't think it should be optional whether to just do the data versus the measure piece. Now, how you get to the measure reliability whether it ought to be this idea of split half versus signal to noise ratio, I'm not sure yet.

I like having both in the toolkit. I don't think I would require the measure developers to report on both, but I do think I would require measure

developers to Matt's point earlier to report for everyone on measure reliability. I don't think that ought to be optional.

Whether they need to do both the data element reliability and measure reliability, I'm not sure yet, but I know that if you're going to present a measure to be evaluated, you need to say something about measure reliability.

And whether that's going to be the signal to noise ratio or split half technique, we can talk about that, but there ought to be some measure of reliability.

(David Behrens): Then we can give developers the, you know enough guidance that they choose an alternative and justify it. Thank you. And you're up again and then Karen and then we should go up to public comment.

Male: So I just want to point out a difference, now if you calculate so called test retest, I mean we mostly have, right? You get one score for the whole measure. If you apply the formula set with a binomial (inaudible) for each facility.

So now if you have 4,000 facility, you have 4,000 reliability score. But when you calculate test retest or split half test, typically you see, you get 1 per 4,000, so they are different.

Male: You present to people what the median reliability is and then the interquartile range. So you do have a number that you come up with, that median number, the signal.

Male: Right, right, so just based -- so yes, it's kind of different, right, but you don't want to use individual, you won't get (inaudible) yes.

Male: Absolutely, absolutely.

(David Behrens): Do we go to public comment? Yes?

Female: Please.

(David Behrens): Do I -- there's an operator who will cast for public comment. (Kathy) please.

Operator: At this time if you would like to make a comment, please press star then the number one.

Karen Johnson: (Kathy) have we received any comments yet?

Operator: No, ma'am, there are no public comments at this time.

Karen Johnson: OK, we're just going to wrap up and if you -- if anyone has comments, let us know. Thank you. We just magically got ahead of schedule by two minutes wow, because we did leave plenty of time for public comment.

We won't worry so much about the other questions on the next slide because we can get to that offline, but I think we did pretty well. I think not exactly completely clear, but we're getting there I think. And I've had a blast. I hope you guys have.

(David Behrens): There's an assumption in your statement that we're complete clear.

Karen Johnson: I am assuming that. I think we have a half hour schedule, is that right, a half hour scheduled for lunch. We'll meet again at 1:15. Are we doing -- yes.

Female: Yes, we are.

Female: Please stay in your seats. We do need to -- sorry -- we do need to take turns. So we have an official hat and we're going to ask you to pull a number out and into the mic announce your name and the number that you receive so we can document what your term is. And we'll just go around.

(David Behrens): You're either two years or three years and it's not a sentence. It's a term.

Male: And what metaphorical hat are those of us on the phone picking from?

Karen Johnson: We'll go off -- we'll just do random assignment for the ones on the phone.

(Jennifer Perloff): This is Jennifer Perloff and I'm holding the number three.

Karen Johnson: I'm sorry, everyone. Could we have just have silence for just a moment. We do want it to be in the recording what the terms are. So we're going to ask

everybody to start reading and so we can make sure we can hear everyone. Thank you.

Larry Glance: Larry Glance, two.

Steve Horner: Steve Horner, two.

Mike Stoto: Mike Stoto, three.

Susan White: Susan White, two.

John Bott: John Bott, three.

Female: Sorry, Christie, could you go ahead and say your name before us?

(Christie Teigland): Christie Teigland, three.

Female: Thank you.

(Eugene Nuccio): Gene Nuccio, two.

Matt Austin: Matt Austin, two.

Lacy Fabian: Lacy Fabian, two.

Male: I'd like to know the reliability of this methodology. And I'm here for three.

(David Behrens): We'll figure it out in three years. Here we go.

(Ron Walters): (Ron Walters), two.

Female: Perfect, and then we'll let everyone who's on the phone --

Female: Oh.

Male: She said she'll be back in a little while.

Female: OK, we'll have Marybeth select one when she gets back and everybody on the phone, we'll randomly select a selection for you. All right, thank you, everyone.

Female: And don't forget that at the end of your term whether it's two or three, you have the option to sign up for another term. So let me put that out there just in case.

Female: With that, please everyone go get lunch and we'll be back in 30 minutes. Thank you.

(BREAK)

(David Behrens): OK, why don't we get started? We're obviously joking a bit about all the easy stuff ahead of us, so there's plenty of deep water here in the discussion validity and we have a pretty large chunk of time in which to do it. (Andrew) is going to do a walking through of slides and questions for us.

But I think just in terms of preface and overview reliability and validity are the two core things that we're asked to assess in the forms that we're doing and although we may continue to tweak this and come up with some other things, our chances are these will still remain on the list for discussion.

There have been questions under the validity label I think, but then under reliability, that would have been this term or that term, what constitutes recent evidence?

So I think this discussion, with the joking aside would be as deep and challenging as the one we had this morning and over a period of time we probably won't get everything completely settled here as well, but it will be different.

And just my own view of this is that the questions about validity really take on an important different dimension in the context in which we're working.

I know in some other technical areas, when we talk about reliability, the way to do it is sort of this same level parallel similar depth of complexity concepts, but in the healthcare quality measure world I think particularly if we think about the subset of outcome measures, there's a different set of questions

about the extent to which a number that we look at, not only is it accurate or fairly reflective of that thing it purports to measure, but then it's another level.

Does this number say something meaningful about an underlying quality of care? One of the most significant quotes from (inaudible) that I remember after probably (inaudible) and actually having the chance to be with him and talk to him in Ann Arbor, in one of his early books that people tend to overlook the concept, he says outcomes measure quality indirectly and in his thinking which we don't necessarily have to accept, it turns out that the main locus of quality is in the processes.

It's in what's done or not done and treatments that are given, not given and yes, you can use outcomes and they're an important part of his typology.

But when you look at outcomes, you have to ask a deeper set of questions about the extent to which movement of a number in that domain reflects quality that is sort of elsewhere. You look backwards through outcomes, to look at what was actually done or not done.

Now, that may or may not help us in our discussion, but I still find the discussions of validity sort of uniquely challenging and interesting, because it's not just about the technical performances and numbers, it's about what do the numbers mean and do they say something important to us about things that are sort of living underneath them? So anyway.

Male: Yes, thanks, (David) for that introduction and we'll get to some of that discussion pretty specifically. What we want to do here is try to kind of finalize how we're thinking about validity conceptually, it's pretty similar to what we did with reliability. And as we did with the reliability, please feel free to just jump in if you have questions at any time or comments.

So to sort of go over, we talked a bit about this on our most recent monthly call. Right now, our kind of formal -- somewhat formal conceptual definition of validity is you know, broadly construed, the correctness of measurement or the extent to which one can draw correct conclusions about a particular attribute, attribute of the measured entity, based on the results of a measure

and then sort of more informally, you could say the extent to which a measure assesses what it intends to measure.

And these are kind of reflected at both the data element and measure score validity level. The data element level, we think about the correctness of the data elements as compared to an authoritative source, so what your -- you know, those specific measure elements that you're gathering, are those things true, I guess you could say, are they correct? Are they correctly reflecting what you meant to collect about that piece of data?

And then the measure score level are the conclusions that you're drawing about quality specifically, sort of (David's) point correct based on the measure score that is, does a higher score on the quality measure reflect higher quality.

And then we have the same kind of little four by four box here as we had with the reliability where you have at the data element level, you're kind of talking about accuracy there and at the performance measure score correct conclusions about performance.

And one question we could ask is whether as with reliability, you could maybe fill in those other boxes in the same way whether you could talk about drawing correct conclusions about performance at the data element level or accuracy at the performance score level.

Just to talk a little bit about some of the reflections you had on the email chain and in our last discussion. As (David), again, mentioned, many spoke about the need for some additional detail on what is meant by what the measure intends to measure.

As (David) said, measures assess quality of care indirectly and can vary in the degree to which those measure results reflect underlying care quality of care. And said that -- you know, some of you said that we would really like to have more clarity and specificity from measure developers and stewards.

On the quality of care dimension that that measure is intended to reflect when you say what it's measuring, what it intends to measure, what exactly did you

intend to measure? Is that really quality? Is it something like utilization? Is it something else?

And to the extent that you are trying to measure some dimension of quality, to what degree is your measure really getting at that, because of -- as we've said, there is no perfect measure but to what degree are you reflecting that dimension of quality that you had intended to -- yes, (Mike).

Male: Sure.

(Eugene Nuccio): Something's gnawing on me based on the other committee I'm on. Prevention and population health and we were reviewing a measure about dental care and the question was about whether or not kids got sealants which are recommended by everybody. It's no question kids should get sealants.

The issue -- but the way they chose to measure it was to say did they get a sealant applied on a particular teeth -- tooth in the measurement year as essentially a proxy of did they get the sealant but, of course, if they had gotten the sealant the year before they wouldn't have as a success.

And as that was pointed out, they said -- this is exactly what we want to measure. Whether or not they got a sealant on that tooth in that year. And it strikes that that's just the justification, of course, for the limitation of the data they had -- that they had available.

I guess I think -- and you can't answer that question by statistics. I mean that really is, it's a measurement specification issue but it's also a validity issue I think. And I get that this question is it measuring what you're supposed to. I'm not sure whether that's in our realm or not and we can't -- generally we can't assess those kind of things from the information we get.

Male: I mean you might hope that they specify exclusions for a scenario like that and say whether -- you know, that the physician or the -- indicated that they had an existing thing or--

Male: Right. There was a discussion about that but even the -- even the basic question is, is this -- what were you intending to measure was…

Male: Sure. Yes.

Male: Quick question. Why do you say you can't judge that based on the information we get? It seems like you could if you get -- I mean you got information and you made a judgment but it doesn't appear to have validity.

(Eugene Nuccio): I got it from that prevention and population health committee but on the measurement committee, we don't get that…

Male: Because we get asked about face validity and that sounds like a face validity…

(Eugene Nuccio): Yes, but we get asked about face validity and -- maybe -- my experience is that somebody said and we ran it through a (Ranpal) and they said it was – has face validity but they don't actually discuss it in a substantive way. Is that…

Female: Yes.

Male: Yes.

Male: (Andrew), when we talk about comparing those numbers results to an authoritative source, data source, I guess the question becomes what exactly, who defines that and how do you define what is an authoritative source and I think this was pertinent for one of my -- kind of measures that I reviewed last cycle because they apparently went ahead and compared something and I didn't like whatever they were comparing, you know.

So that's -- I mean I think that's we need to put some time also. We need to discuss that issue as to how we define what should be the comparator.

Male: Yes. And I know there -- in some types of measures there are sort of standard I guess accepted things like for example, for (FUEs) and something -- a claims-based measure typically the authoritative source is considered to be the patient's medical record.

And I would -- let's say, I'm not a clinician but I assume that there are often mistakes in the medical record as well but it's -- I guess it's as close to the

truth as you might be able to get plausibly. So that's usually considered the authoritative source.

I think in things like instrument-based measure, we often expect there to be validation of the -- of the instrument. And I'm not an expert by any means on how that happens, but I think they usually expect there to be some -- I don't really know the methodologies by which these scales are validated but to show some internal consistency or some consistency over time.

Demonstrating that that you -- the question you asked and that the answer you got was actually reflected some truth about what you were trying to get at. I think that's probably, there's -- I don't think we're ever going to get away from some subjectivity or some amount of error there.

But -- I don't know, it may be something that we could offer some guidance on, what are good authoritative sources of truth or closest to it as we can get. I don't know if anybody else has any thoughts about that.

Male: Go ahead, (Paul).

(Paul Kurlansky): I think that the -- this is -- sort of gets to issues brought up earlier in the day and that is -- perhaps one of the limitations of the data that we actually receive in this committee is we don't get the whole picture and sometimes I think it might help to address the issue of validity if we got the whole picture.

In other words, somebody -- it seems to me theoretically conceivable. Somebody could come up with a reliable measure and present data as to it being valid for such and such, a general topic. And yet if you actually look back at what they're driving at with this quality measure, it doesn't really fit it.

And I -- so some -- it's like the (inaudible) people do study reviews or if you sometimes read the study protocol and then read the article, you find two different things, right? They changed -- they changed everything.

The original study protocol was designed to answer this question and they couldn't answer that question so they present the data on what they could answer. And that doesn't really address the issue.

And so the validity of the measure may be very much tied to the question that's being asked. And so we need to know the questions being asked in order to answer the question of validity.

Male: To follow up on that. There's a general belief it seems amongst many developers and users like CMS that claims-based data are -- give us more reliable and real information than various surveys and other clinician assessments.

But they fail to in many cases recognize that claims data are for fee-for-service and don't represent the patient population.

And so when you start looking at utilization outcomes, in particular like hospital use, we found that hospital use using in the home health world dropped suddenly from 28 percent or 28 -- about 28 percent to 16 percent in one year. And has maintained 16 percent over that time period.

And that was simply the change in the data source when we went from (Oasis) that included all home healthcare patients to claims data which only included fee-for-service, we suddenly had a 12 percent improvement in hospitalization utilization.

So the data source is critical with regard to validity unless we want to make some sort of descriptive change in the -- in the name of the measure, that is acute care hospitalization for fee-for-service patients.

Male: Thanks. I have (Joe) on the phone and then (Christie).

(Joseph Kunisch): Yes, hi. Just a quick comment on the validity and the materials. I'm -- I remember on the first round, we were told to only use the materials in the application that the measure developers supplied.

And I found myself spending a lot of time actually looking for the articles and so forth that supported what they were proposing, and then even in the next round again I spend a lot of time going outside and doing my own

investigation to say, is what they're actually saying here accurate? What -- can I find this in the literature?

But if I took a gist from the application, I mean I'd think they -- from my perspective they would fail on a lot of points just because they didn't give me the right information that I needed to make a decision, is this actually valid data that they're getting.

And then around the performance and I get very concerned especially from the clinician perspective, that when you say this is showing better performance -- well, if you're going to define what performance is, you better have the evidence to show that if I do this, it's actually going to impact the patient care in a positive way.

There has to be enough supporting literature for that outcome to make that. Otherwise, all that you're measuring is the change one direction or the other. So that's it.

(David Behrens): (Christie)?

(Christie Teigland): Yes. Just to piggyback on an earlier comment about the claims data. There -- I think you really have to be consistent in the sources of data that you used. So one of the measures that I evaluated required HbA1c lab values and it -- and also it said you can either use claims data or if you don't have access to claims data, you can use medical record review.

Those are two very different things. Lab values are not typically available on claims data unless you have your in-house lab or something and you get -- put it, gets on the claim. Less than 10 percent usually 10 to 20 percent of the time, you might have a lab value in a claim.

The problem with the way the measure was defined is that it wasn't just looking at whether the person got a test, which is always in a claim, but did they have a lab value. If they didn't have a lab value, they are considered not in compliance. Their HbA1c was not in compliance. So they were -- you know, they were dinged by that.

And so that's what really made me reject that because that -- no, it doesn't mean that the person's out of compliance with the -- with the HbA1c level, it means it's not in the claim.

And so maybe organizations that were using medical actual record data were going to look a whole lot better than most of the plans who then were using claims data. So the source of data, really you can't mix up sources of data like that is the lesson there.

Female: Just a quick point. I think one of the questions I've struggled with validity is content expertise. So a lot of us are methodologists, we don't have content expertise and I wondered about the overarching committees and where they come in -- on the validity question particularly. They will have knowledge that we don't have now.

Male: I was actually just having the same thought. That much of this is -- does seem to be a little less methodological and more conceptual and clinical in many instances.

And we do get to the -- that sort of question of the -- whether you can, in fact, impact the outcome, but in the evidence section, which is not something that this panel reviews, we ask that the developer show -- give us some information to show that the outcome that they're measuring can in fact be impacted by at least one process, structure or intervention of some sort.

We used to only ask for just a rationale showing that that might be case. We're now asking for a little bit more substantiation of that, but really just some -- maybe some literature showing that the outcome can be influenced by some action that's taken.

But it is -- it does seem a little challenging or problematic that this committee is sort of separated from that question. And also from these questions of reflecting the actual kind of clinical quality that's being delivered.

It's less of a methodological question than it is many -- in many ways conceptual or clinical. And if -- and especially when you get into these questions of exclusions or risk adjustment factors that are appropriate or not

appropriate, again, you can look at the risk adjustment methodology but the question of whether some factor or not should be -- should have been included or some exclusion, should or should not have been included is really kind of a clinical questions that goes to the clinical expertise

And the Standing Committees do address that. They have the opportunity but that we're sort of taking that away from them in some ways with this process. They do have a chance to check on it but we've removed the emphasis a little bit from that.

And I don't think that they're going to not address it because they -- they're very much interested in talking about that. So they will take their prerogative to discuss it and address it. But it -- I understand that it would create some difficulties for you guys to be sort of separated from that question of evidence and its connection to the rule--.

Female: I think the only thing I was going to -- and you got to it I think, (Andrew). That's exactly why we allowed the Standing Committees to overturn your rating because many of you are clinicians but not all of you.

And even if you are, you might not be conversant in this one particular thing that the measure is about. So that is exactly why committees can overturn. It is a little hard.

As a matter of fact, it's probably impossible going into this initially we were kind of thinking, well, you guys need to be paying attention to the method was it a sound method and were the results reasonable and decent sample size and all, that kind of stuff.

But if you have information and clinical background and experience about some of these other things, of course, that goes into your thinking as well. So there is no bright hard line that we could reasonably draw there.

Female: Yes, I just wanted to second the points being raised about the crux of the issue to me, which is going back to how these measures are actually going to be used for quality improvement and accountability. And I'm jumping ahead to look at the questions later on in the day, but it's relevant now with validity.

I think validity is the critical -- I mean that's the critical factor. If a measure is not valid, if it's not measuring what we think it's going to measure, it doesn't matter if it's reliable. It doesn't matter if it's easy to measure, it's useless.

And I think to the point of the clinical expertise versus the methods expertise in taking a look at that validity -- I think it comes into play a little bit with where we were going with reliability, that there are different ways to go after those methods.

And one of the things especially from the psych background is thinking about things like cognitive interviewing and actually talking with clinicians and actually going on the ground so to speak to ask and figure out conduct semi-structured interviews on whether or not this measure is actually useful or meaningful or seems to be assessing what it is much more beyond face validity.

And those are the kinds of methods -- I don't see that often. And we don't see them that often because they're hard. And they take a lot of money, they take a lot of time. They're difficult to do well.

And a lot of those things frankly measure developers don't have the luxury of a lot of times with various systems in place and what measures are needed right now which gets them to kind of a whole different issue but I do think it's really important conversation today though going back to considering how validity can be a first consideration because it is, in my opinion, the critical factor for these -- for these measures and how they can actually achieve their objective.

Male: As a clinician, I can attest to the fact that I think it's sometimes very difficult to know how much impact clinicians and hospitals and providers have on performance -- on quality of care.

Part of that is because we frequently don't really have a good sense of what best practices are as much as -- the evidence base on what works is actually in many cases quite limited, especially when it comes to very complex care protocols.

You know, there are certain types of interventions where it's quite clear, but very rarely are clinically interventions just binary. There's a whole lot of things that put together is usually a bundle of care. Whether those bundles actually impact quality or not is hard to measure.

So at the end of the day, what people really look at I think when they're looking to see if there's -- if a performance measure reflects quality of care is to see whether or not there's variability between performance -- between providers essentially.

So if you have variability then it's sort of like prima facie evidence that there's -- there's a reason why those providers have differences in performance levels, it's because they're different.

And the problem with that is that it all has to do with your risk adjustments. So if your case mix adjustment is not very good, then maybe a lot of that variability that you're seeing is just basically because of differences in case mix and not because differences in provider quality.

But at the end of the day, I think -- at least when I've evaluated a lot of these measures, much of whether or not it reflects quality has to do at least in the measure developer's eyes, it has to do with whether or not there's variability in provider performance. And that's what they usually hang their hat on.

And then that's why it's so critically important I think for this group to look at the risk adjustment and determine whether the risk adjustment is valid or not because if it's not, because if it's not, then that variability, a lot of it may go away.

Male: So just sort of another question is whether there are any assumptions about validity that should be questions within the ou forms what we're asking our developers or facets of validity that we're missing.

Again, there was some discussion on the email trail after our last call that talking about the assumption that to be valid, a measure must be reliable and

that maybe we shouldn't betying that quite as closely and should be thinking about them a little bit -- more separately and distinctly since I -- yes.

(Eugene Nuccio): I know that these terms are used differently in different fields and some people do say that. But it seems to me that there's a lot of value in treating validity and reliability as two separate concepts.

And that in fact, we currently do treat them as two separate concepts, it's 2A and 2B, whatever they are. Whatever the numbers are. And to say one is a pre-condition for the other, I think just complicates things in an unnecessary way.

Male: OK. I guess I would observe that even though that assumption that is listed here may be floating in our work somewhere, the structure of the valuations do not strictly require this. Is that correct? That we work through our assessment and we judge reliability and then we keep going and we assess validity. It's not as if we get -- if we fail a reliability, we stop.

Male: That's -- it used to be when we were doing this -- these valuations at the -- just with the Standing Committee, we would and we would do them in sequence. You know, we would be around the table like this and vote at the time. And if the measure did not pass reliability, we stop.

Male: That's not the same. That says that if it doesn't pass reliability, it's not going to be approved as a measure.

Male: Right, right.

Male: That doesn't mean that it's not valid.

(David Behrens): I'm wondering -- not to derail this any, if we could just flip back to the last slide, I thought you were going to say a little more about the last thing and then -- it's a little dear to my heart which is why I care about it.

The signal to noise issue typically comes up in reliability and I think it's got a reasonably clear conceptual understanding even if the metrics escape me

sometimes. But if we're looking at a number is that a true number or is it random of noise. That's sort of getting in reliability territory.

In a couple other contexts, I've had to -- in a group like that, I've had to make the argument that there's also an appropriate way of bringing up the issue signal to noise under validity but it's a different line of thinking, but it basically says if I'm looking at number that's either high or low, to what extent does that number reflect underlying quality of care.

And that's just a different kind of question. But what it gets at is this sort of (inaudible) distinction between process and outcome.

And most of the measures that we valuate can be classified as either one or the other. And often the argument about validity regardless of whether it's process or outcome depends on what the developers can bring forward about the known link between the two.

And it is a process measure stands on some ground if it has been shown to link to outcome or predict the outcome, the doing of something or the not doing of something is better outcome.

But in parallel, the standing of an outcome measure may depend on the extent to which it is influenced in any known way by some aspect of process, which sort of gets a little bit of the issue of case mix and risk adjustment.

So, for example we might want to look at the measure five-year survival is a -- as a quality of care measure or performance measure for cancer centers or for oncologists. But in order to do that, if I'm going to judge validity of that measure, I might want to see some evidence that it is something that they do or don't do that affects the outcome as opposed to histology stage, comorbidity, et cetera.

So I do appreciate it, I just ultimately want to say thank you for putting the signal to noise thing because I think it's a line of discussion that can have an appropriate home under validity as well as under reliability. It's just that it's a different kind of discussion. OK. OK.

Female: When I was reading kind of back and forth the email chain and, again, especially what you were when you brought up that point.

Kind of the way that I -- the point that I thought you were making with that and maybe I was wrong but this idea of signal, I felt like you were suggesting that there's more in signal than just quality of care. So did I read too much into your comments?

(David Behrens): No, I think I could -- I could go with that wording, but now, we're going to get into detail about what I mean by that anyway. You know, if we look at a number, we could use this five year survival, and you -- say you're comparing two -- let's stick with readmission, let's take two hospitals.

I got a high number and a low number. My question here in the validity is to what extent does the difference between those two numbers reflect quality of care, different quality of care provided by those hospitals versus something else.

Now, when we're in the reliability territory that “else” is essentially random error. It’s sample size -- it's -- doesn’t have substantive meaning itself. But under validity, I think it does have substantive meaning.

Its case mix, its risk, its bias, its distortion, its -- but are -- but all these things that I have in mind are things other than quality of care.

So that's why when I use the signal to noise, now I'm saying if I look at a number and it’s higher or lower, if I look at two numbers and they're different, to what extent is the highness or lowness are different a reflection of underlying quality of care versus something else.

And in this case, the (else) is more than just random error. The (else is meaningful, substantive stuff that often has to be adjusted away.

Male: And does the question of say -- you know, adequate risk adjustment address what you’re--

(David Behrens): Well, it's in -- it's in that territory but I -- it even can go deep in the net because I think sometimes we have measures brought forward to us where the original case to establish that the number reflects quality of care is very weak. Before you even get to risk adjustment.

Now, sometimes that case is plausible and then you have to deal with risk adjustment. Sometimes I'm still looking for the evidence that says this is a good quality measure before we worry about risk adjustment.

Male: Let me see if I -- this is how I understand it, too. So we're talking about a measure outcome, usually a process type thing tied to an outcome and that's the cleanest one that doctors like and so on.

That I give these treatments and outcomes are improved that's what we see all the time. Besides all this stuff that you mentioned as far as risk adjustment, think of all the other processes that are going on in the care of that patient which are not reflected by the measure adjustment and are subject to all sorts of random variation.

Nursing care, home healthcare, you name it. Everything else that that patient undergoes while they're being measured with this outcome but for this measure and -- yes.

It's very complicated and that's the problem that (Eric) points out all the time is -- it is one of many contributing processes to the outcome that is usually tied back to said to be affected and justify the validity of that measure. But it's just one of many, many, many…

(David Behrens): And that's where my challenge to measures of developers is not that they're -- must show us that there is no (else) at all.

It's just that there's a sufficient signal that it can give us something meaningful given -- first of all, you adjust away all the (else) that you can and there's still some residual (else). I just want to see some evidence that the residual -- the signal is strong relative to the residual (else).

Male: Just to sort of take a different perspective if you're looking at it from say the consumer's perspective, it all -- all of that stuff I kind of don't care how we get there, that all matters, you're still getting a better or a worse outcome than another hospital and that's what I'm interested in, in seeing that I can say that this hospital has better outcomes than that one and that's where I want to go.

I'm not, you could, it could be all those kinds of things, I don't really care which processes you tie it back to. Can I draw a meaningful conclusion that you're producing better outcomes than another place?

Male: And we talked about…

(David Behrens): Dr. (Kurlansky) are you producing them or are you just accidentally--

(Paul Kurlansky): Yes, accidentally. And we talked about uses earlier in the morning, the reliability part but, I mean, attribute, just go to any discussion about attribution and you've got everything there.

Male: I think David raised an issue which is even more profound than risk adjustment and attribution and that is and particularly when you mentioned readmission, that's a perfect example because even if you have it well risk-adjusted and perfectly attributed, the question is is it a measure of quality because if you're looking at congestive heart failure, it's inversely related to mortality.

So, it is a measure of utilization but is it a measure of quality. And that is I think that's another layer of thought that has to go into this. Even if you have a reliable measure, it's highly measurable and you can risk-adjust for it and you can attribute it appropriately which is that's saying a lot, but even if you have all of that it still may not be a valid measure of quality.

(David Behrens): Thank you. Jennifer?

(Jennifer Perloff): I think we're on the edge of a bit of a slippery slope. I think that'a sort of where you're going here. How these measures get used is beyond our control. As methodologists, we can't bake in the appropriate use of these measures all the time.

So, in federal programs, we see colon cancer screening as a measure of lung surgery quality. Obviously, that's not a good alignment and that's an extreme example.

But, I mean, in this model of what variants do we have control over, there may be things like social determinants of health that today you don't have control over. But if you change the construct, you could.

So, potential influence, actual influence, that's the slippery slope that I feel you can go down and so, I like to retreat to science and avoid some of those complicated questions. So, I just wanted to point out the edge of the processes.

(David Behrens): Yes. Just to try to cast a little bit of a walking path on the slippery slope, the reason I sort of favor this signal-to-noise is I think within our scope people could tell us what's the observed data at least in the context we currently have about the relationship between things that can be done by the entity being measured and the outcome and then we can form a judgment on that basis. It's either, does it appear strong or not strong.

And then if the external environment changes or a measure-user wishes to use it sort of outside the scope, I'd agree that that is outside our scope. But I think for example, what often comes to mind is let's say the evidence base is for measures of a clinical trial in which an intervention produced 10 percent improvement in mortality.

And the clinical trial was perfectly well done, it's published in New England Journal, it's great. But what you're comparing is doing it all the time versus doing none of a time. And now in quality measurement you'd be comparing two doctors or two hospitals where the difference is 10 percent, the difference is 5 percent.

Now, my rough mathematics says the complete doing and not doing produced 10 percent difference. The 10 percent difference in doing is going to produce a 1 percent difference in mortality.

OK. Now, is that a strong enough signal to justify either the outcome or the process in this context? But that, I think, is within the realm of science still but slippery slope. OK. We had John I think and then we have Jack on the phone.

John Bott: Back to a comment that David made of looking at signal-to-noise to inform risk adjustment and how it's working. So, I just recently had a conversation with John Adams and he had an interesting conundrum he was wrestling aloud with me about.

He was noting that – I hope to get to characterize this correctly – he was noting that a problem he's seeing is that the risk adjustment is working, right? You're letting all this bias come through. You're sort of amplifying the variation that exists, your manufacturing variation.

So, in turn what happens, you can see a horrible risk adjustment and you're driving up the signal, the apparent signal that's coming through. So, he noted that that was an issue with signals, noise as it relates to risk adjustment in trying to evaluate it. But that's a good insight from John – another John.

(David Behrens): Jack on the phone.

Jack Needleman: Yes. OK. Not, on that point with John. It depends whether you expect the risk adjustment to narrow variation or to expand variation where there are real quality differences that are masked by safer or less safer patients that are being treated. So, it can cut either way.

The conversation we've been having strikes me as I heard the slippery slope and one of the questions it raises is where the Standing Committee's work begins and our work ends. So, is this really a measure of quality? At one level we're asked to judge that in the sense of validity. Is it measuring what they claim it's measuring?

But the broader issue of it is that quality measure feels to me a slide into the Standing Committee issue. We also have the issue of sort of precedent and how that establishes where we go in terms of assessing validity and that comes up in the risk adjustment modeling quite explicitly.

And on things like readmission where as was noted whether or not the patients comes back into the hospital depends not just on what happened in the hospital, but also what resources are available in the community, what unmeasured severity of health literacy or capacity to self-care by the patients none of which are particularly well-measured in our measures.

And when we look at the socio-demographic risk adjustment, the measures that are routinely available in the CMS data are frankly lousy for socio-demographic measurement in terms of some of these other issues of community resources and patient capacity. And that's reflected in the marginal adjustments that may extend in terms of the risk adjustment estimates.

But we've by precedent at this point pretty much established an acceptance of a risk adjustment model that's based on the (HCCs) and minimal or no socio-demographic adjustment even when there are strong community factors out there, and have sort of kicked over to the Standing Committee to say beyond this is the adjustment good enough.

So, we've got this issue of what are our precedents. To what extent are they binding on us or should be informative to us about what we do. And where does our assessment of validity as it is measuring what it thinks it's measuring and then the Standing Committee's begin.

Male: I'd just like to pick-up on that and propose a specific response that might – see if people agree so we can maybe make a little progress -- and that would be that the purview of this committee when it comes to the form, because there are some kick out points in the form where if they fail, they fail.

Its reliability is low, it's out. The validity is low, it's out, that this particular aspect of validity, so once we've move beyond the risk adjustment is good, the reliability is good, the validity seems on the face of it have potential to be sound.

But you may have a serious question as a reviewer on this panel about whether this is really a valid measure, that that would not be a reason to reject

sending it to committee, that we would still send it to committee because that would be the purview of the committee, although we could express the question or even the concern that the committee could consider.

That would avoid kind of a double indemnity to the developer and at the same time allow us the chance to speak our mind on it, but not have it be a basis for rejection.

(David Behrens): There were a few nods. Did we pass something? Thank you for framing it, Jack.

Male: Yes. We'll have to figure out a way to sort of operationalize that but…

Male: I had one more question about that one. So, that is specifically at the measure level. Do you feel the same way about the data level?

Male: I think so. Let's see, are we going to talk about, yes, we got a question here about does it make it any sense to talk about, think about accuracy of the measure score or the correctness of conclusions about data elements. Is that getting to the same thing?

Male: No. No.

Male: Not really. Yes. And we talked -- there's another – I thought there was something in here maybe it's in the criteria section about whether we should continue to accept data element validity as essentially counting towards data element reliability. Right now, we do allow that. We'll talk about that later. All right. Let's get moving with that.

Male: We have Karen with a comment, a question on the phone.

Karen Joynt Maddox: Yes. I just wanted to sort of echo the prior comment about the fact that this probably crosses between the methodology and the content expertise which is largely concentrated in some of the content-specific committees.

If you think about something like, I don't know, say, hospital-acquired infection, if you want your measure to tell you where the highest rate of infections are and where, therefore, we should direct resources to try to

improve infections, you actually don't want to risk adjust at all. You want to know where the highest rate of infection is.

And we got into this a little bit in the readmissions/admissions Standing Committee around a measure of, I don't know, it's local readmission or admission, I can't even remember what the measure was but it was not risk-adjusted.

And it really depends on the intent of use, whether or not you want to explain away what some of the differences are from or if you want to use the measure to try to figure out where we need to be directing resources.

And because we can't as the methodology committee necessarily know all the context around how a measure -- I understand we're also not supposed to think about use but it's even hard to gather the context sometimes as others have pointed out.

I think it is a slippery slope. And I don't think it's an avoidable one but it may require actually some back and forth with some of those content committees or doing what I think David said, which was to say here are our concerns.

But perhaps the content committee can best address some of how this might be approved or maybe even place some parameters around what appropriate context sort of might be for thinking about something like that.

(David Behrens): Yes. Jennifer is tagged up but no Jennifer to speak. I guess that makes it Paul.

Paul Kurlansky: I think you raised an extremely important point and that is what constitutes a "good risk adjustment model." It depends upon what it is you're trying to accomplish.

So, if you want to drive up your C-statistic for example in the world that I know, in cardiac surgery, if you want to drive up your C-statistic, you would include pre-operative factors, intraoperative factors and post-operative complications.

And you can get it up there very high. But if what you're trying to find out is are the patients that you operated on comparable, then you would only include preoperative risk factors.

And this is where I think we have to at least get out on to the slippery slope a little bit because as a methodological issue, a methodology which included all of those things and came up with very C-statistic would be methodologically sound but it wouldn't answer the question that you're trying to answer.

(David Behrens): All right. Let's go back to Andrew and I'll pick up Jennifer when she's back.

Andrew Lyzenga: On that issue, the risk adjustment for social determinants that you guys worked on is very clear about what you can adjust for, things that exist at the start of care and so on. And I've always had the impression that applies across the board but I'm not sure that it's actually stated for the NQF.

Male: I believe it is.

Andrew Lyzenga: Yes.

Male: So, I think we want the clinical risk factors as well to be present at the start of care.

Andrew Lyzenga: So, that would address that.

Female: Although to be fair it's stated definitely in the report that you helped write or what we call our SDS report. I don't know if it's in our guidance. It may not be like written in that document that you're working with.

(David Behrens): All right. So, to sort of get to some of what we were talking about earlier, should we add the following ideas to our current definition which I guess we'll go back to that to remind you what it was.

The extent to which one can draw correct conclusions about a particular attribute based on the results of the measure. The questions are do we want to add that idea of the extent to which a measure assesses what it intends to measure or adequately distinguishing between good and poor quality.

To some degree I might argue that that's inherent in that question of, that phrase about a particular attribute that you may be -- whatever you're trying to measure is it the quality or utilization or whatever, that particular attribute, you could define it in whatever way you like. But we may want to be more explicit about good and poor quality.

Male: I have to say I don't know what that means.

(David Behrens): The definition as it stands?

Male: Yes. I guess that one, the one about the correctness of measurement.

(David Behrens): Yes.

Male: I just don't know what it means.

(David Behrens): It depends, I guess, again, on what you mean by your particular attribute and that sort of depends on the measure and how they've defined it.

Male: Could someone provide an example? I think it would be very helpful to have some examples.

(David Behrens): Right. So, I guess maybe talking about the readmissions, your attribute, I mean, it would be a whole set of attributes really.

The degree to which you provided quality of care in the hospital, given appropriate and high quality discharge instructions, many would broaden it to say are we providing adequate social services or transportation. The whole whatever – that sort of concept of what an institution might do to affect readmissions, yes, some are much narrow than that.

Male: But that strikes me as totally within the realm of the subject matter experts as opposed to a methodological, I mean, that's fine but I can't imagine how we could evaluate that from the work we do.

Male: These are, I think, both examples at least, what I was trying to suggest, we could speak to them and give opinions but they wouldn't be a basis for not forwarding to the substantive committee. If these are the only concerns that

this committee had about reliability and validity, it would still go forward is what I was suggesting.

But we should at least if you can render an opinion, you should render it or ask a question. I mean, I think these are sort of the same thing, aren't they? Is there a real difference between these two bullets?

Female: I guess it sort of goes back to Lacy Fabian, we're going back to the point though that if we're saying we would give opinions about this before it goes to another committee, one of the things I feel like that other committeea should just make the decision first because there's no point in us doing the prior stuff.

Male: A lot of work.

Female: Like if there…

Male: Like a journal editor that says you know what, I'm not going to accept this article no matter how the reviewer thinks about it.

Female: Yes. I'm not trying to be cheeky, I guess it's not I don't see the point in doing any of the reviews.

Male: (Eugene), to go back to the other question that you had on there about distinguishing between good and poor quality providers let's call it, as has been said numerous times by numerous people, a provider does not have total control over the outcome or the utilization.

So, are we making too broad a statement that – can we distinguish between good and poor quality that's under the control of the provider or the provider's protocols? So, I think we need to condition – set a condition on this. It sort of gets to David's point about (Plato's) what's the real world and all of that sort of stuff.

And then not to be a lawyer but is the word correct conclusion the right term or I would use the term appropriate conclusion, because you can draw all kinds of conclusions. I don't want to judge correctness but appropriateness I would be willing to judge.

(David Behrens): OK. I think we've got Joe on the phone and then Karen as she had her tag up and…

(Joseph Kunisch): Yes. I just wanted to kind of support that comment about providing that feedback to the measure developer but maybe not saying a no pass on going further on endorsing or going back to the committee that should make that decision as how it relates to the actual quality of care provided.

I think maybe this goes back to what the valuation form ends up looking like how we change that because I know in my last one I did that, I put this is what I thought from a more of a clinical perspective.

If you're saying this is supposed to improve or show improved performance, what is the actual performance outcome that you're improving? If you're going to make that statement, you need to back that up.

So, I like that idea of being able to provide that because I might see something from more of a clinical or implementation side of things that I can give that feedback at least back to the main committee for them to help decide whether to endorse this quality measure or not.

(David Behrens): Let me just ask a question. I supposed I'm turning to Andrew and Karen. When the developers submit the materials for measure endorsement, I assume somewhere in there they're asked to make some sort of declaration about the intent of the measurement.

Let me just give you an example that if I bring forward a measure readmission since we've been talking about it, I could probably do three or four things.

I could on one that hand at the surface, I could tell you this is a measure of readmission. No other inferences. No underlying assumptions. It's a measure of readmission. Boom. Done. And then we could ask is it a measure of readmission. OK.

Or I could come in and I say this on its face is a measure of readmission but what I'm really trying to measure is quality of care, about performance by the hospital as opposed to something else. I could make that declaration.

And then it would seem like we could say OK, what's your evidence that this number reflects that concept. Or I could deeper and I'm playing off what you said a few minutes ago, I could say I think this measure reflects the quality of the discharge instruction.

Then we could say OK, what's your evidence that that number reflects that content. OK. So, do the measure developers bring forward that kind of declaration about the concept that the measure is designed to reflect?

(Andrew Lyzenga): A little bit. Not maybe quite so explicitly as you might like. We have a section where we ask them to provide the rationale for the measurement and that's kind of a free form. We didn't give too much of instructions there.

I don't in fact recall correctly but usually that's the place where they can say this is the reason we're measuring this and that usually entails some justification related to why this is related to quality, why we want to measure this, why do we think this is meaningful to measure and then to some degree the evidence section, that connection of the outcome to some sort of process. But it might be worth focusing that a little bit to try to get out a little bit more of what you're saying.

(David Behrens): Yes, because what I'm trying to find is this boundary between what's us and what's the Standing Committee.

And I could say if the developers have to declare that this number is a reflection of this concept, I think it's reasonably at least most of the time within our scope to say what's the evidence that the measure reflects the concept whether it's the right concept or a good concept or the best concept and maybe could be something that flips to the Standing Committee. Could we work with something like that?

(Andrew Lyzenga): Well, I think that's a really important question. My experience in the prevention of (inaudible) health committee is that it's often not there. You're right, they have an opportunity to do it and sometimes you can infer what they're thinking. I call it the theory of the measure from the important section.

But that story I told of the dental field was the clear failure to do that and it was only after they were pressed did they say well, that is what we're trying to measure which seems inconsistent with the important data.

So, I think that we really should ask for that we really should ask for that explicit statement that how is this supposed to work, how is this – exactly want measure – how is this measuring what dimension of quality. We can't really assess validity without that.

(David Behrens): What do you want us to conclude about the provider or the quality of care or something like that?

Male: I mean, it's a little bit analogous to study without a hypothesis, right? And so, I think it should be an explicit requirement. What exactly is the intention of this measurement? What is it intended to measure? What quality metric or outcome are you specifically targeting?

Male: (inaudible).

(Andrew Lyzenga): I would fully support maybe having measure developers being more clear on the construct that they're trying to measure and maybe to sort of maybe reinforce another point is maybe our group shouldn't be the one to necessarily put that meaningfulness or value on whether that construct is really -- (inaudible) readmission. I mean, to me we should be assessing is it actually measuring the admission.

I mean, I think we've tried to put a light on to readmission to what that means as a measure of utilization or a measure of quality and I'm not sure if that's the space we should go into or more the Standing Committee should make that judgment call. But I would support the idea of having (inaudible).

Male: Excuse me.

Female: Yes.

Male: (inaudible).

Male: Yes. I just want to make a quick follow-up. I totally agree and what makes me nervous as a non-clinician is I feel like when I'm making that validity, I feel like validity for us as the methodology group should be a recommendation.

It shouldn't be a go-no go because of all of those issues, because of I don't know that we're equipped to measure whether it's meeting the construct even if the construct is defined. Simpler ones maybe but…

(David Behrens): OK. We have Jack and then Sam both on the phone.

Jack Needleman: Yes. So, this discussion of readmissions I think illustrates again the issue of how far are we going.

My recollection of the history of the readmissions measure is CMS' original concept was to look at the claims data and identify avoidable or unnecessary readmissions. And they were going to stop paying for those.

And then they realized they couldn't do it out a claims data. So, we wind up with a higher than expected level of readmissions comparative measure on the assumption that if your readmission rate is higher than expected a lot of those patients, those are the ones that are probably unnecessary or avoidable or influenced by the quality of care got in the hospital.

And that in turn led to risk adjustment, a model to get the expectation right. And that would be what we would see and we'd be looking at the expectation, the risk adjustment model and so forth.

But the question of whether the readmission is in some sense the higher level, higher than expected readmission is actually a good measure of quality feels to me to fall over to the Standing Committee that's dealing with that because it requires a lot of more substantive information about what causes readmissions, when they're avoidable, how they're influenced by things that we don't expect the hospital to be able to control and so forth.

So, we can look at the goal which is higher than expected readmissions look like a problem with quality. We can look at are they actually measuring

readmissions, are they capturing them in the measure, that's the measure level. Does the risk adjustment seem appropriate given the limitations of data and is it good enough and that's the score assessment.

But is the underlying measure legitimately one that measures quality and, therefore, should be included in measure set feels to me to flow over to the Standing Committee side.

(David Behrens): Sam?

Sam Simon: Sure. And I would certainly endorse what Jack just said about Standing Committee really being in the best position to assess whether the measure is meeting its intent.

But I would say also representing a measure developer perspective that having the opportunity to explicitly state the intent if the measure while it should be done by all measure developers it's probably going to be really helpful to standardize this and I'd support that.

Karen Johnson: And I just want to make sure, I think I'm understanding that you're suggesting this idea of measuring what you intend to measure, that sort of thing that you really do feel like that you can opine on that but that shouldn't be a make or break for validity; that would just be input to the committee and they would go forward with it.

I want to make sure though that other facets that we asked you to think about would you agree that there might be some things that are within your purview that would make you say no. So, let me give you example.

Data element validation with sensitivity specificity and maybe they didn't quite do it right or their values were not good at all. I want to make sure that you're not saying we want to opine on that but not until you know as well. Is my question to you clear?

(David Behrens): And Matt gets to answer.

(Matt Austin): This is opinion of one, I mean, to me that still feels like something we would want to make a judgment on because that feels more methodological than conceptual to me.

Male: Yes. I can agree. I think going back to (Sherrie's) earlier point like (inaudible) data and you know that (inaudible) something that they necessarily know that that's not going to be too (inaudible) we should be making that judgment.

(David Behrens): Lacy?

Lacy Fabian: I want to make sure I'm clear, too, on the distinction to me seems like the classic like when you're in graduate school the statistical significance versus the clinical significance. It's not like I keep kind of coming back to that like we can comment in the methods panel about the statistical significance.

But I'm trying to think of a good example and some of the measures that we've reviewed thus far, I wouldn't feel comfortable making the judgment about the statistical significance and only offering an opinion on the clinical or like the meaningfulness of it if I could tell that it's not meaningful like if you could statistically show that 10 is different than 12 but can tell that has no like basis for reality, being worthwhile to distinguish and show a valid measure, I don't think I could just say yes, that's statistically valid but I really think somebody down the line needs to pay attention to whether it's meaningful if it doesn't look like it.

(David Behrens): Yes. These are process points. I think since I have looked a couple slides ahead, there's a slide with some -- called additional questions that speaks directly to this point about what's a meaningful difference.

Do you want to get into that now or do you want to wait because I think that opens the door to a whole interesting line of discussion.

Male: I'm sure we can talk about that. Yes, we asked about meaningful differences as part of the (assessing threats) to validity.

Now, the question I think we wanted to ask really was how this relates to reliability and we certainly talked about that a little bit with the signal-to-noise and the ability to differentiate between measured entities, is it redundant and are we adding value with that question about meaningful differences.

That's sort of what we are thinking here but it seems like there are some broader questions, too, but I don't know if this is something that you want to opine on. When we ask about that, we're asking for some statistics. We're asking for them to demonstrate that there are statistically significant differences among the measured entities.

And we mean meaningful in that sense statistically I think is how we have been thinking about it. But here is a different meaning to meaningful that we might think about.

Male: I got a little confused there with the meaning of meaningful. But I actually just want to make a comment back to what Lacy was saying and it made me think that I might have sounded like I was suggesting people not express the view that it's not valid. I would think it's perfectly fine for someone on this committee to say I don't think it's valid, I don't think it's meaningful.

It's just I was suggesting it not be a basis for not going forward that that opinion would then carry forward to the sort of parent committee if you will, the substantive committee and they would review it.

That avoids double indemnity. I mean, I think it's not fair to developers to give (inaudible) two shots at opinion rejecting but you should say what you think. I mean, you think it' not valid, say it. I just wouldn't use that as a reason to not go forward.

Male: I think there's a methodological construct for what Lacy is pointing out. It's one thing for example, for me to opine that readmission is not a valid measure of quality which is really beyond the purview that I could put that opinion in but it's beyond the purview of this committee.

On the other hand, there's a concept of (inaudible) sides. And so, if you have a large enough number and if you look at large data sets, you get significant differences but, right, but are they meaningful.

And you could get a one percent difference that has functionally no meaning but is very validly derived that statistically significantly different and I would think we should be able to make an opinion about that.

I think that opinion then has to go to a higher level or a different level committee to see what they're going to do with it. But if the methodology creates a situation where the effect side is so small I think that might be a valid reason for us to question it.

Male: The question about predictive validity and those other (inaudible) I forget the different kinds but oftentimes we'll see measures saying well, this is a measure of access to health care and we see that people with low income have lower access. We think, therefore, it's a valid measure.

That's my predictive validity. That's a different (manner) of what is right but I don't know if that's what this is getting at, that kind of consideration.

Male: I don't know. I was, again, we had a fine discussion at lunch that the analogy I think of and check with Lacy to make sure, see if I'm on the right track. In the world of clinical research and particularly in some of the domains of patient-reported outcomes, there's the acronym MCID, Minimum Clinically Important Difference.

And there's a whole psychometric industry that goes underneath but basically it's just how much improvement should there be in order for a patient to judge it to be meaningful or worthwhile having gone through a treatment.

I use this regularly in the context of the spine surgery (inaudible) you have a whole bunch of patient-reported outcomes. And based on published literature, each one of them has an established MCID value.

How much improvement do you need to see in the 0-10 (inaudible) in order for patients (inaudible) consider it's worthwhile. And then we build a

performance metric that says in what fraction of your patients do you achieve MCID and we risk adjust it up, down and sideways and we use it for quality improvement.

So, I'm thinking about your question for the (inaudible) context that as we look at performance measures brought forward to us and then we sort of are under this validity section, if there are sort of room in that discussion for something analogous to MCID meaning are the differences that this measure can detect meaningful in the substantive sense not just statistically significant but are they meaningful. Would anybody care about them?

And part of the problem comes from this question of huge sample size. When you work in the hospital domain, you're working (inaudible) measures, you can do a whole bunch of analysis with 5,000 sample size and you can produce statistical significance or significant differences where the actual movement of the number is to my subjective taste sometimes is really small.

But then I don't know quite where to take that. In the world of patient-reported outcomes and MCID, I've got some ground to stand on. I can say that the change for this patient from here to here was or was not above this threshold.

I can look at the performance of surgeons or groups and I can say on aggregate for the patient (inaudible) they treat risk adjusted to some of them more frequently achieved this improvement in their patient than others and we tend to take that as a measure. But I don't know where that corresponding ground to stand in is here for many of these measures.

Male: Is that a reliability issue?

Male: It's not really because it's not just about purely statistics. It's about in the case of MCID, it's the judgment of patients of whether the improvement is actually meaningful to them like would you do surgery again given the improvements you've had. It's a substantive judgment.

It has – I don't know what (inaudible) – it's got meat to it. It's not purely a statistical term. There are statistical methods to arrive at it but it has judgment. It's about values. It has that component to it.

Female: I think another example to make it more concrete as you think about like a utilization measure, if you can statistically, if you have a large sample size and you can statistically show differences in cost of $100, is any hospital anywhere going to care that something was $100 cheaper like is that going to be worthwhile or meaningful.

Female: I would argue.

Female: Probably not.

(David Behrens): Yes.

Male: When you put this in front of us is that this may not even be what you meant by meaningful.

Male: It's not really but this is a conversation worth having.

Male: Yes. So, I think that example that you made of the MCID is a really great one. And I think that it's particularly applicable when it's not a binary outcome although a lot of the outcomes that we are looking at – many of them not (inaudible) all utilization but are binary.

And so, in those cases we don't really get to make a, I mean, it's you live, you die, you have a complication, you don't have a complication, it's usually meaningful if you have the bad outcome.

And I would say I think we kind of have consensus here. I think most of us agree that it probably should not be a must-pass for this committee, that we should get to comment on it but at the end of the day we are not providing the content expertise; the Standing Committees are.

So, when it comes to things like importance to measure and whether it's appropriate, whether the measure construct measures what it's supposed to

measure, I think that's probably more in the domain of the Standing Committee.

Male: Yes. I think we're hearing that and not actually hearing much disagreement. It's just, again, are there any sort of lines that we can draw that establish a boundary or even I think and is this a question we should even be thinking about given the constraints you just mentioned.

Is this something on which we could pass an opinion legitimately to the Standing Committee and I could (inaudible). I got Bijan and (Gene).

(Zhenqiu Lin): Yes, Paul, I think going on the, I mean, I'm somewhat confused. So, going back to David at this point, so are we saying that essentially we would not – this committed would not reject any measures that would potentially (inaudible) not the validity question was not satisfactorily answered and (inaudible) feel that way.

Is that what we are saying that we cannot reject any measure on the basis of validity? All we are going to do is probably recommending or saying this is what we think. Is that what we are going to do?

Male: I wasn't meaning to reach into all of validity (inaudible) I think not all validity…

(Zhenqiu Lin): But don't you think this is the meat of validity, I mean, (inaudible).

Male: Well, maybe we should, I mean…

Male: That's what maybe a discussion we can add, where can we draw that line.

Male: Yes. Yes. Where is that line? Where is that line? Yes.

Male: What's the appropriate decision for this committee to make and what isn't.

Karen Joynt Maddox: Things like missing data and others elements of validity may be very (inaudible) methodologically (inaudible) I would argue.

(David Behrens): And Mike?

Male: To whoever was that who spoke I couldn't recognize your voice.

Male: Karen.

Male: But, yes, Karen. The missing data on exclusions is something I think that we can comment on as it relates to validity from a methodological perspective. So, that was going to be my first point.

My second point was that if with renewal measures that we may be seeing coming down the road especially process measures that are all at 98 percent with a difference of 95-100, validity, they're useless; they're not meaningful. OK. So, I think we could probably measure, we could comment on that from a methodological perspective.

Male: I should note that with this committee, unlikely to look at any process measures. We're almost exclusively looking at outcomes unless there is composite process measure (inaudible).

Male: There will be with IMPAQ. Just saying, which is already at 90 something percent.

(David Behrens): Mike and then I have Joe on the phone.

(Michael Stoto): So, I've been trying to think back to the measures that we've reviewed both in this one and in the Standing Committee that I'm on and what's presented in terms of statistical information in the validity section. Sometimes you see they report an analysis of a panel that it says (face) validity.

Sometimes they do construct validity which is the example I couldn't think of the name before. Sometimes they do concurrent validity where they say it's related to other measures of the same concept. Sometimes they say predictive validity.

But (inaudible) at least they have statistical measures. Sometimes even for face validity they have some kind of statistical measure. I mean, it strikes me that those things are all within our purview but that we really need to

understand the theory of the measure before we can judge whether or not the data they provide really are appropriate.

(David Behrens): That's great. Joe?

(Joseph Kunisch): Yes. Just a quick comment because I think those are some valid points as far as some members expressed maybe not having the expertise or knowledge to even make comments and I think that's very appropriate.

But there may be times that some of us do and I think it's providing that feedback but I also wanted to remind everybody that there are other venues of providing that type of feedback also.

So, I would definitely say it shouldn't be the go or no go deciding factor but just another way of providing the feedback to the Standing Committee but then, of course, they always have those open for public comment and typically before CMS adopts some into their programs, they release some in their rulings where you can comment on if you feel the concept is actually measuring quality. So, I don't think we're doing, asking to do anything out of the – that's unusual I guess I should say.

Male: And this maybe is a question for NQF staff. So, my understanding, the formation of the Scientific Methods Panel was really help address issues where folks on the Standing Committees maybe feel uncomfortable with trying to understand some of the more sort of mathematical or statistical tests.

So, to me it's like what is the gap that we're trying to fill or what are we really trying to achieve here. It's more of a question I'm not sure I have the answer to that but what comes to mind is this idea of I would think what might be helpful to Standing Committee might be almost sort of like a memo that says we reviewed the risk adjustment models, they have high C-statics, you should feel comfortable with that. You should feel comfortable with the data element validity (was high).

And more almost instead of a go or no go but really sort of maybe inform that committee like we did review it from a statistical standpoint we're

comfortable, it's really sort of up to all of you to figure out if readmissions is really something meaningful and useful to measure.

Karen Johnson: So, to some extent that's somewhat baked into the process I think because the way it's set up now if you guys see a statistical problem so we're not talking about the clinical side or some of the stuff that there's a risk adjustment model and the calibration thing, you just don't think it's a good model or whatever or bad sensitivity, whatever.

If you think that's the case then we send it back to the developers. So, in that scenario, the Standing Committee never even sees it. But otherwise, if you push it forward, you are saying that we think statistically everything looks pretty good.

But then the idea particularly with validity is this committee will bringing that clinical side and they really do need to think about some of these things that you guys might not even think of asking, right, because you're wearing a method hat not necessarily a clinical hat.

So, I think that's kind of in reality what's happening or at least what we meant to happen with the process. I don't know if that makes you feel any better or not but…

Male: I'm just a little confused about the process. When somebody submits a measure, what happens? In other words, does it come here to see if it's methodologically sound and then go to committee to see if it's – I see.

So, if somebody submits a measure which is sort of like ridiculous but methodologically sound, it would still come here and then get rejected down the road, is that how it works? So, basically, we're the first round gatekeepers?

Karen Johnson: You are round one the way we have it set up. So, what could happen is you guys spend much of time and say the (stat) look fantastic and the Standing Committee looks at it and says your evidence doesn't really back up what you need to do and we're going (inaudible) on evidence.

So, if we did it the other way around, I guess that's possible but we have to really think about timing and all that sort of thing.

And even if we did flip it, we would still have to take it back to the Standing Committee at some point, I think probably for these meaningful differences things, because there's -- in a way, it's chicken and egg, right? There's kind of no reason for them to be thinking about clinically meaningful differences if they did their testing incorrectly.

(Paul Kurlansky): I mean, I sort of put on my editor's hat here and wonder if there -- if there is a possibility to do what certain journalista do, unfortunately mine doesn't, but where the manuscript comes and the editors look at it, and 50 percent of them get thrown out right there, without even going to review.

And then, the other ones OK, and what I get, I mean, the statistical side is, the editors feel this is potentially publishable, we can -- can you look at the methodology.

So, if I'm wondering if there isn't a sort of a step first which could be made that this measure just has no -- it ain't going anywhere. And so -- and there's no reason for this committee or other committees to waste time on, it's gone it, going and goes back to developers.

If it looks like, oh, potentially, then, OK, make sure it's methodologically sound. And then, oh, OK, make a decision on it. I'm just throwing it out there.

Female: OK. I completely agree with Paul, I think it's that workflow step, but also links up with what Mike was saying about the theory, because if we don't know how the measure is intended to be used, we don't know which statistics, like, we can't make -- I don't feel comfortable making statements about the statistics, if we don't know how the measure's supposed to be used.

So, there's some kind of either having the form that states that theory or having the workflow change to have that initial knockout of, does this even make sense before we look at this.

Male: I -- can I give an example of where I think it's important, it's going back to the readmissions issue. If you're thinking about readmissions as a way of measuring the quality of the hospital per se, you may want to adjust for whether or not the patient has somebody at home to help them, on the other hand, if you -- if you're doing it for an ACO where you just care about whether or not they get continuity of care, you would want to adjust at all.

And so, but you can't make that judgment unless you really understood the whole way supposed to work in that context.

Male: I just OK. OK. I'm slightly different.

Male: OK. I was going to note that we are -- we're working on a proposal to incorporate something we're calling graduated measure review.

Where we would allow for the submission of concepts and measures that are specified to a certain degree but not yet tested, and we would bring those to our committees, and with either with staff, we give feedbacks sort of along the way, and that might serve as the steps sort of a gatekeeper function where we would provide feedback.

And once it got to the point where it was submitted for panel's review probably would have gone through some vetting of that sort.

(David Behrens): OK. I want to get back in a slightly different angle to this meaningful difference thing, and I --the difference angle is -- I'm now trying to bring in a concept that's either very much like are actually is the concept of a power calculation. And I think this is more squarely within the domain and more statistical.

And the question I'd be asking of a measure and therefore the developer would be, given the parameters you've given us, including -- say, a table of samples and methods, how small a difference can you detect with this measure?

Now, before we are meaningful, we are about what's meaningful to patients or this is more purely statistical. But I would really like to know that about

measures, because if somebody comes through -- and then, again, like, all of this, and I'm thinking one I had in one of our earlier sets.

It was a physician performance measure that did a lot of the pretesting on a huge sample of thousands of physicians, and then they came back and said that this will work on a sample of twenty five patients per physician.

I'm sorry, they did it on where they had thousands of patients and physicians and they said that this would work just fine if you have 20 plus. And I said, well, how do you that?

But let's even assume that they were correct, but somehow there's a number that's still would close here, that if I've got a sample per physician of, say, several thousand patients, I might be able to therefore detect a quite small difference between physicians and declare it at least to be statistically significant.

With a sample of 25, I think it's different. Now, maybe this just falls under reliability and we've already talked about it, but somehow I don't think we had it quite that squarely in front of us, it's somehow -- I would like to see that given to us and then I'd like to see us be able to judge is that calculation and is that claim valid? Now, is that -- tell me if we've already been over this ground already, but I don't know that we have.

Female: I'm going to say, I fully agree, I had one measure where they guessed at a minimum sample size. And I would have liked to more -- and here, because they're possible, they could have had done something that was minimum. And so, I agree that I think it's an important to mention that developers can share with us.

Female: I wholeheartedly agree, and I think that's a question that we should -- we should actually add, because when they're testing the measure, they actually have all the statistics we need to make a stab at it, right?

It's not, I mean, it's not like that situation I used to be in the (grants) where I was sort of saying to the researcher, but you have to give me something, I can't just spin the wheel. So, yes, I think -- I think the inclusion of that is very

important. And at the same time, I think we can give that our own face validity check, right?

If it takes three years to get enough sample size to be able to detect a difference then it's not a good six-month measure and then I've seen measures go through in the early days that were very much like that, were based on three years of data, but it's OK, we used it on per year or something like that. But I think you're right, I mean, I think it's important.

Male: I totally agree that that's important, I just happen to think it's really reliability rather than validity, but, yes,

(David Behrens): I was watching our clock going and actually given the remote location I now live, I have to be on a 5:00 flight, so I thought I had -- because I'm in my little window, if I don't say it.

Although, I will trying to be back (inaudible) I'm going to dash to the airport and rejoin you by phone. It was in here somewhere, it sounds like that, there's at least some support for trying to make that easier, so, yes.

Male: I think that's mostly what we wanted to address. I mean, we kind of talked about some of these issues, I mean, we've gotten into some places we didn't anticipate. I personally kind of like this idea of this panel serving more of an advisory function on validity, sort of, like, there's some questions, some questions, some elements of validity.

And it may take some work on our part, we need some more advice from you guys on where we draw that line, and exactly how we distinguish between those questions that are within this committee's purview to say no go and which ones we want to just provide more advice to the Standing Committees.

Those have been really helpful. I mean, if you want to talk about this question of thinking about accuracy and of the measure score or correctness of conclusions, about validity.

(Paul Kurlansky): OK. Can I just say one thing we can do over time is track the the percentage of times where this committee sends a negative statement about validity, as if

it would have so negative it would have rejected it, and see how often that gets approved.

You know, it may never happen, which means it's working, and if it happens a lot, then maybe we need to revisit it and say, you know what, the things we think are getting through that they're not -- they're not somehow hearing us or, maybe we're wrong, or, I mean, we should -- we should look at that too. But I don't know, there's something we can track with data over time and experience.

Male: Sure. Any other thoughts on or things -- does anybody wants to add about validity?

Male: Going back to this point, I think how we communicate to them would be very important, I mean and it's like -- I don't know, I mean right now, say for example it's go or no go, maybe in all of these, there will be another (inaudible) laid out, so it has some level of uncertainty.

And then they probably would take it. And if you're right, I mean, often times comments will not be read, but that way some sort of sort of something that they this is what, this can be found.

Female: And part of the trick here too is, it's -- get that it's very different just kind of thing where your purview ends and then -- and kind of how far you need to go, this takes us back a little bit to our discussion this morning about process, right?

Because -- and especially the form that you use because even how clearly a Standing Committee who may not have much in the way of statistical or any that kind of expertise would be able to interpret your stuff.

That comes into play too. So, if it's -- so I'm not being very articulate, and I apologize. But all of these things that we're talking about are not divorced from each other, they all kind of interact.

Male: And I should just mention, we hear you about the request to hear more about that question of the quality construct too and let's try to incorporate that into submissions.

Female: Just real quick, because I know it's 3:00 and we all really want a break. But does anybody have a good -- I shouldn't even ask this question, anybody have a good definition of a quality construct that you could help us find?

I mean, I kind of know in my head or maybe we don't get that formal, we just say, tell us what you're trying to do with this measure.

Male: Well, one thing I do in my class is I ask the students which of the six (inaudible) is this trying to get out. And that narrows it down, and then within that, you can -- you can work from there.

Male: I also feel like it should be pretty brief and it's just sort of my initial reaction, but it shouldn't be paragraphs to describe what you're trying to measure, that it should be you should be able to say in at most one sentence or two.

Male: Yes.

Female: Yes.

(David Behrens): No, I think that's fair, but a good measure should be able to do that or measure developers and say, this is a measure of X (inaudible) concept is this, and it's a safety measure, it's a patient experience measure and off you go.

Male: Whether or not you -- whether or not (inaudible) skills?

Male: In my template, I think it's perfectly straightforward. Yes.

(David Behrens): All right. Are we overlooking anybody on the phone before we break? OK, 10 to 15 minutes again.

Male: Yes, yes, yes. OK. That was a meaningful tap. OK. All right. So, we're going to -- we're going to have our final session. Through Final -- well, no more breaks.

Female: Yes. No more breaks.

Male: No more breaks, we're going straight through to the end. And I think there's people are still on the phone with us. And David's going to call in when he gets through security at the airport. For now we're going to move to measure evaluation criteria, a discussion, and Karen's going to lead, I'll facilitate.

Karen Johnson: Sounds good. And real quickly and apologies if the team has already asked you, but does anybody plan on leaving this room before 5:00, even if we're not done? OK. What time are we planning?

Female: About 4:00 pm.

Karen Johnson: 4:00? OK. 4:00? OK. OK. All right. 4:00. OK.

Male: Or you should pack in -- pack in this next 45 minutes…

Karen Johnson: So, we're going to -- I'll talk really, really fast.

Male: The most important questions coming this next 45 minutes. All right.

Karen Johnson: All right. This piece may not be quite so fun, but it -- we're going to -- we've talked about -- we talked about these things already a little bit and if we don't come to final consensus, this is -- this is OK. But I think the way I want to do this is just talk about the first question first, let's have our discussion and get through it.

So, very quickly, should we be asking about validity before reliability? And that -- a knee jerk is yes. The implications for that is that we would have to potentially change our voting, some things like that, change our algorithms?

You know, I think about what all has to change maybe a little bit. So, does anybody have any -- is it just, yes, we should or is there any hesitation?

And the other thing, let me tell you, in terms of specifications, right now for reliability, we say think about the specifications, are they complete, are they unambiguous those are the foundations of the measure. So think about that

and then think about testing. If we did validity first before reliability, you might not have a chance to think about sets, so.

Male: So, you asked if there were any reservations, I see some -- I see maybe a couple going up. But I -- if I had a reservation, it would -- it wouldn't be -- I wouldn't want to express it as a vote against starting with validity, because earlier I suggested the possibility that we could start from the measure generally and then deconstruct down from there and not necessarily thinking in terms like validity and reliability.

But just break it down from top-down and to some extent, that would tend to bring validity up before reliability, I think.

So, I'm generally in favor, but I think it would have to be understood or qualified in some way that because some aspects of validity -- because to some extent, validity depends upon reliability, you can't do a complete job on validity until you look at reliability.

Or you might say, well, I happened to peek it, I happened to sort of sneak a peek at the reliability enough to know that it's not valid because it's not reliable. Because reliability can be a necessary, but not sufficient precondition for validity, so that -- but I think we can overcome that.

Male: I'm going to say this, although it won't makes me the most popular person in the room. So, we -- I told Karen we just kind of decided to become consultation on validity and absolute on reliability. Now, I don't mean that as literal as it sounds, but kind of that's what we did.

And so, I'm concerned that -- about what's going to happen to reliability if we base how we handle reliability based on what we feel about validity. In other words, we're going to probably be -- again, consultative, suggestive, and so on, on validity.

And while I think validity should come before reliability, actually reliability just got priority in our previous discussions. All things considered. So, I agree, I -- my mind goes to validity first. But that's not what we just kind of talked about.

Male: Well come back to that. And Karen on the phone first, and then (Gene).

Karen Johnson: Yes, so, I guess I think one thing that I've learned from our conversations today is that, reliability and validity are not two monolithic standalone powers of statistical trustworthiness.

And instead, what we've identified is sort of a subset of things within both of those. There's face validity, there is use validity. The different reliability constructs sort of stood out as being probably that we need to separate out elements versus measures.

And so, instead of thinking of these two things a sort of two mountains that we need to cross, thinking about sort of a more holistic evaluation of a measure, I think anything without face validity can stop there. If it's a terrible idea, it should just stop, I think that's fine.

But I'm not sure some of the finer points of validity and reliability are really quite as linear as we've made them out to be. I think everything needs to be assessed, and I don't know that there really should be a stop point for either one before the other necessarily.

Male: Previously, we spoke about asking the developers to put the intent or the purpose, goal of the of the measure in there, I think with that information along with the measure specifications, numerator, denominator, exclusions kind of thing, that could give us -- give me, anyway, a sufficient understanding of whether or not this measure is a -- in any way useful and meaningful and so on and so forth.

But then I tend to be -- again, personally, I tend to be a constructionist as supposed though with destructionist. So I find it easier to think about it as looking at the reliability of the source information, i.e. the data element, and then the way that the metric is constructed to see if we -- if we can in fact create differences, meaningful discriminators between provider groups, and then move on to the validity thing.

So, I think with some of the suggestions that we've made, I would argue that we should have that front end stuff to help us understand what the measure's driving at. But then continue with a form in the reliability validity out for the details of the validity and moving on.

Male: So, I can kind of -- I'm not sure about this. So, we talk about -- we -- I guess supposed to provide advisory roles on validity. So we studied down right there, so we already know I mean, in that role, right?

Karen Johnson: Yes. So, let me -- let me make sure that we're all clear on the -- what we're calling the advisory role for validity. And I'll add some NQF flavor in there.

It's -- I hear very well, and as a matter of fact, you're already doing this with some aspects of risk adjustment, right? You're playing an advisory role, but there are certain things in validity that you can be more than advisors, you can be very specific and definite.

So, when we talk about this advisory role, what -- at least where I'm sitting in and where we would be very comfortable with at NQF is that there are still -- there could be very valid statistical methodological reasons that you guys would say no to a measure. And you should feel comfortable and empowered to do that.

There are some other aspects that you wouldn't feel comfortable, the meaningfulness for example is one, things that factors in risk adjustment models whether you should or shouldn't include a particular SDF factor, that sort of thing.

It might be another -- I'm sure there's plenty of others they would be more advisory in that perspective. So, is that the collective understanding by everybody? Let's at least make sure that every, like, I see a couple nods, and not so many nods, OK. Does anybody have a different understanding? Maybe if you put it like that.

Male: I mean, I'm concerned about, like, different people may apply that criteria and a little bit different maybe you would cause some inconsistency across, you know…

Karen Johnson: And I think that's the part of the squishiness that we all have to just live with. And maybe, I mean, at some point, we could probably start tracking it to see if there's anything that we can be more systematic about maybe, it's I think it might be a little early to do that now.

Male: So, how about -- how about if -- rather than try and get a a full consensus right now, we write down within the validity category the things that this committee would have purview to reject? You know, for example risk adjustment. Well, let's say -- well, face validity that would be one example.

And just excuse me, be a basis for just not going forward, I mean, this is just not measuring what they say it's measuring, and you can see that on the face of it. Or risk adjustment, because I think there's some feeling maybe expressed by (Ron), but others too that we may have just given away the ranch.

And I didn't necessarily have that impression, just that we were making, I mean, setting it up so that decisions that were more appropriately handled by the relevant committee are sent to that committee with our -- with our input.

But that doesn't necessarily mean if this risk adjustment there's no risk adjustment and there should be, right? That would be we could say don't go forward. So, I think -- I think there are bases for validity to that this committee would have, at least that's how I've been thinking of it.

But it would just be the one to -- well, it's is a two percentage point difference maybe the data show that you can actually differentiate people based upon two percent, but we don't know if that's meaningful, we suspect it isn't, but we don't want to reject it, and we want -- maybe we write these things down. And that would make it more workable, and not give away the ranch. Larry?

(Larry): So, Karen, I think you started off the discussion by asking which one should we do first, the reliability or the validity.

And I -- and it's -- I really like the point that Karen (inaudible) made is that this is really kind of a big picture thing and it's kind of holistic. But I'm still

going to be a little linear, OK? I think we ought to start first by looking at the data elements, OK?

The data element reliability, if the data elements the risk factors that you use in your risk adjustment models aren't valid data points, then the measure falls apart at that point. So, that'd be the first thing.

The second thing that I would look at is the reliability of the risk measure itself. If there is no signal, there's really no point at looking at the risk adjustment, there's no signal, you can't discriminate between providers, so what are we doing here? So, that would be the second thing.

And then the third thing is I would look at the risk adjustment, because if you're seeing a signal, but if you're doing a lousy job with the risk adjustment, and you're not taking into account the fact that different providers have very different case mixes, maybe a lot of that signal or maybe all of that signal is because of differences in case mix as opposed to differences in provider, true provider performance.

So, I would say start with the data elements and then looking at -- so, then look at the reliability or the signal to noise ratio for the measure itself. And then lastly, look at the risk adjustment itself.

Now, in terms of the validity, and I think, David, you were talking about that. Yes, I think we ought to be looking at the kind of like the face validity of the measure, I mean, does this measure feel right?

But I think at the same time, we should recognize that we are not the content experts on this. So, it should not be a must pass for the measure. I think our expertise is sort of more in the methodological realm. But if something is absolutely crazy, of course.

Male: So, maybe what prompted this question around which one to do first? Are there problems with the way we've been doing it or is it -- I guess I'm just trying to sort of understand that…

Karen Johnson: Yes, just what -- why are we even asking? I think it was -- it goes back to our assumptions when we talked about assumptions of reliability and validity and to be valid a measure must be reliable. It's late in the day, and I might be getting it backwards.

But anyway, not everybody agrees with the way that we said that. And part of that statement I think probably made us do reliability first and then validity. So, I think that was part of it.

I don't -- I don't know at the end of the day, because they're both must pass, right? So, you have to have both in order to end up with endorsement. So, at the end of the day, it may not matter too much, and if it doesn't matter too much, then we probably would want to just keep it as it is.

If there's a very compelling reason to flip it, then I think that's what we want to know from you. And what I'm talking about here too is our process in terms of how we work through things in the committees as well, right? We always talk about vote on reliability, then go to the next one.

You guys could certainly, there's nothing stopping you with your form. You guys going to question whatever on the validity section and doing that part and then going there's nothing stopping you guys from doing it in however order you want. So, does that help?

(Paul Kurlansky): Just a question on that. So, let's say we go to the form and validity is a non-pass, do we not fill out the reliability portion?

Karen Johnson: That is where you guys are at a little bit of a disadvantage, because there is no stop for you guys, right?

Because it could very well be, because there's a non-consensus across panel members, that you're feeling that it doesn't pass validity, reliability, so therefore, I don't want to do validity we may you might change your mind once you hear your colleagues or vice versa.

So, right now, and there's a little bit of -- a little bit of inefficiency in there, because we're asking you to evaluate even though you feel pretty comfortable

that it may have missed the other. It's not completely wasted effort, because your feedback can be very important to developers and the public, et cetera. Yes.

Male: I'm wondering why we're worried about this. Is it because of we have so much work that has to be done, we're trying to cut back on it, I mean?

Karen Johnson: I think it's more philosophical than anything. So, if it's -- if it's -- if we shouldn't be worrying about it, then let's not worry about it.

Male: Yes. I mean, I'm thinking we might -- we have something to say, we're only advisory in the end anyway. Why not just say it?

Male: So, I think going back to Daniel's point, honestly, I'm just getting used to the form, I mean, I know that the form has a lot of issues and lot of it, but then over the last two cycles, I mean, I'm just getting used to it, and I know (inaudible) and I know what to look, I mean, I'm just getting to know rather, and I don't really want to again have some sort of drastic change and again so that's sort of my personal opinion.

Karen Johnson: I think you -- is yours up for a question? OK. So let's put that one bed, let's talk about minor updates to the flow of the algorithm. And I think what I want to do is just make a statement and you guys tell me if you have any objections to it.

What we've learned and we sort of missed before but our algorithm have a little bit of messiness in it, and we -- when we built the form, it really came into play, because we were trying to make that form work with the algorithm.

So, let me make a statement, and you tell me what you think about it. You see a measure, and well, and this is for validity, right? This is on the validity side. Or actually it's about -- it's about -- OK. You see a measure and we asked first, just on the form, and in the flow about floor level for liability, OK? And so, let's say you see a score level reliability that looks great.

You know, no trouble, looks great. But some information, they -- the developers went ahead and told you about data elements reliability. And it

doesn't look so good, OK? The score level, yes, you can distinguish differences, but you can't consistently pull weight.

Would you want that measure to go through because the score level looks great, even though you know the data element doesn't look so good? Or would that -- would you want to say, hey, data element, you know?

Male: We have some opinion.

Karen Johnson: All right.

Male: So, in a content, so I use the instrument, right? So the instrument you get a score, scale score, and the instrument may include a dozen items.

And part of the reason you want to have that to include a dozen items even knowing overall score will be more reliable than one single data item, right? But it cannot -- somewhat inherent there.

(Eugene Nuccio): Yes, I mean, I agree with that, and I also come back to the point that the scores are at the -- at the level of the provider. What's that? Oh. Yes, that the scores -- that the score's really at the level of the provider. If they do it's reliable at that level, I really don't care about the elements.

But I -- but I want to -- but I would like to come back to the suggestion, I forget who made it this morning that our output should be a memo, maybe a structured memo that covers everything in an organized way, saying, here's what we saw as the strengths and weaknesses of this -- of this measure.

And consistent with our jobs has become being advisory to the Standing Committee rather than -- rather than sort of making decisions on our own. I think we can sort of put some of these things in perspective better if we did that. We still may want to go through the algorithm to make sure we're consistent.

Male: Yes, I think going back to (inaudible) your example, I'd be very circumspect, like, we get -- if I get to see that the core level measure is good, but then like not that it seem -- their element levels are not as good.

And when I get -- and I think about the score level (inaudible) sometimes it could be mean or something (inaudible) aggregate level here just might be good, but then you really need to wonder as to what -- where it is coming from, and that would -- that would be some would say I would be worrying about.

Male: And I second that, and, I mean, give an example BMI is a really easy measure that the people compute. But even if I get a good score there, I'd like to know if it's self-report, height and weight, or if someone actually measured height and weight. So, the data elements are important for me.

Male: And I -- and I second that, I think the data element reliability is absolute critical. If you're looking at outcome live or die, it has to be real. If you're looking at the risk factors injection fraction, age, gender, congestive heart failure, these have to be real.

But I do understand the point that you're making is that if the data elements are a composite, so go and you're creating a score out of those data elements, not every one of those data elements necessarily has to be reliable.

So maybe we need to make the distinction between what you were discussing and the data elements that are specifically included for risk adjustment and for a not a composite outcome, but a single outcome.

Karen Johnson: I went to -- along these lines, I reviewed a measure where the input data was a surgical outcome measure or some type, and input data covered half of those surgeries, but half the surgeries were done, because there's registry based measure, and outside of that specialty with no visibility of literally half the surgery.

So, everything was technically correct, but I had to question the end measure, because they wanted to say this is this quality of whatever the surgery was. And so that's a philosophical question, not a technical question, but it seems related to the notion of the data -- the input data, I just wanted to raise that as a dilemma I faced.

(Paul Kurlansky): I mean, I sort of support the opinion that it -- the data elements are not reliable, even if the measure appears to be reliable, it would concern me, because it means that I don't really understand where it's coming from.

If I don't understand where it's coming from, then that means it may be reliable in particular situation where it was where it was presented to me, but it may not be applicable outside of that very limited context, I'd very concerned about it.

Male: I'm looking through the sheet, I'm just making sure I understand what you said. So, score a level rating, question six. You can do that. I mean, wonderful score, and then you go to 10 later on, because you can't avoid getting to 10. And it says the rate -- yes, rating data element.

And choices are moderate if score level testing was not conducted, low if score level testing was not conducted, and then -- or insufficient. That's your three choices, I think.

Michael Stoto: So, I -- that's -- but that -- this is the reason why I proposed this idea or, I mean, or somebody else's idea about writing a memo.

I don't if the BMI was based on self report, I would like to know that too, of course. And I think that that's worth saying for exactly the reasons all you -- all you people mentioned. But I think that the algorithm doesn't necessarily get there.

That we just have to be conscious about these things and lay these things out that's a -- if we find that there's a good reliability at the -- at the score level but not at the element, data element, that's kind of something you need to think about, and understand better you can make a decision about whether or not you endorse the measure.

Male: So you don't think this is, all of you or not just -- not just you, Mike, but you don't think that we could improve the algorithm to make this less prone to individual judgment?

Like, it sounds like something -- maybe I am starting to check up what I'm hearing, see if I am hearing it right, that the general view is, yes, it's important to have the data element information in addition to the score.

And but maybe it's not possible to improve the algorithm. And that because we have to -- you have people exercise their own judgment and do their own investigating on the data element, is that right?

Michael Stoto: No, no. I guess, I think that our contribution as people experience and expertise in the area is to use our judgment. And that the algorithm should support that but it would -- but it shouldn't be getting in the way with that.

Male: Matt, do you have a comment?

(Matt Austin): Yes, if I understand the algorithm correctly, it sounds like if there is measure score reliability testing that it does not require data element reliability testing. And so, to me the question is, if I'm measure developer who has supplied both, are you going to look at those?

But if I'm measure developer who only supplied measure score reliability testing, and didn't give you any information that might be the element reliability testing, I'm going to get passed through.

As a measure developer, my incentive would be to not provide any information about the reliability testing, unless it's both are going to be required.

So, to me it's -- to me that might be, but the things we need to make, are we going to require both, or are we going to just require one? To me, if we were just require one and someone supplies both, then that feels like a little bit of an extra penalty.

Male: It sound like you're a lawyer coaching a witness, right? You know, just answer the question and nothing more.

(Matt Austin): I've watched a lot of legal shows.

Male: Yes. It's a really good point, I mean, I think -- I think -- you're -- you didn't put it this way, but it does seem like given what we're hearing, we do need to specify that if there are -- that if there are complete data elements that contribute to this score, we want to see them.

We want to see the reliability of them. That it's a requirement, it should be in the algorithms. Is that what you're hearing? You're not liking what you're hearing here.

Karen Johnson: You're making it really, really hard on some of our developers, but we hear what you're saying. And to be honest with you, I, number one, I am not surprised, and number two, sort of agree with you. And I think that the trick will be how do we balance what we really love to see versus what's a reasonable ask. Yes, we're…

(David Behrens): Is this the group muscling in on reliability because it fears it's lost validity? So, we have to -- I think we need to maybe sleep on that and come back to it in subsequent discussion, but anywhere -- comment from this point, (Jennifer)?

(Jennifer Perloff): I'm just going to say it -- couldn't a developer make a compelling case why they might not be able to provide?

One, it's not -- they can't provide the item level, because you're claims data example rings true to my heart, it's really hard for me to go get medical records audits. But -- so, if I can provide a compelling case, maybe I get a pass. Even though we can still ask for both, so not to let that drive us away from asking for stuff.

(Paul Kurlansky): Bring up again some points made earlier in the day about the importance of what we do. And I've I hate to be a hard ass, but I mean, it's -- these are measures which are going to affect reimbursement and possibly credentialing and other sorts of things. I don't think it's inappropriate to give the developers a hard time.

Female: Yes, speaking of things near and dear to my heart, and here's one of my -- being on the Standing Committee on Disparities comes in. You know, where

it's particularly hard to validate the data, it's for socioeconomic risk factors, as Jack mentioned earlier.

And my sense is, yes, it's really hard to get this data, but that's often given as an excuse for not testing, not risk adjusting for -- or even testing for risk adjustment, when -- even when there's a compelling case to do that people are living in poverty that might impact the likelihood that they have a bad outcome compared to somebody who's wealthy and healthy.

But we can't test that because the data is too hard to find, the data is too expensive, all of those is that a -- is that a valid reason when the resulting measure then is not adjusted for socioeconomic factors?

And it might be not a good measure of how reliable that is showing you a performance of physicians? If I'm taking care of really, really poor people, all of them compared to you who have all these wealthy healthy people. So, yes, when should we give a pass?

You know, it's just because you can get the data, it's not -- it's not like it's not out there, we know that it is, or even if you use insufficient data if you use readily available census data, and you've heard me say this a hundred times, Karen, but if you use a readily available census data, but it's it's at the block group level, or is that the five digit ZIP code level where you've got rich people and poor people, and so the effects wash out.

So, they say, oh, yes, we tested it, and there's no affect. Income has no effect poverty has no effect. No, you used the wrong data, so it's also about the data that you use. And then, I don't know, do we just excuse that and give them a pass? So, I'm concerned about that.

Female: And Larry, your card is up, but before -- you have to leave at 4:00, right? I want to make sure we get to the next little piece, and you give your spiel, but do you want to finish anything on the criteria real quick?

And we may come back a little bit later. Is that OK with you guys if we kind of skip that last bullet? We might come back to it if anybody wants to. Our

toolkit, and since Larry is leaving at 4:00, we wanted to talk a little bit about what might go into the toolkit, that sort of thing.

But Larry really had suggested, and before you leave, I think it would be really an interesting thing if you just give the panel your idea for the second thing, article in a peer reviewed journal, and just kind of make that case. Is that something -- if you wouldn't mind doing that?

(Larry): So, I think part of what we're doing basically is coming up with a consensus on how to evaluate, how to evaluate measures. And I think clearly that's a really difficult process. And I'm not quite sure we're even halfway there yet. But I think we will get there and I think it's really important.

And when -- I think part of the importance of that is that it really sets the bar for what measure developers will be expected to do, because if we decide how we're going to evaluate measures, then we can essentially communicate that information to the measure developers, and then we'll all be kind of on the same -- in the same place.

And, again, I really don't think we're even halfway there yet, I mean, even for something like the reliability of the data elements, I mean, we could probably spend half a day still talking about that.

Because as much as I am a very strong proponent in the idea that garbage in garbage out, if the elements aren't reliable, I mean there's no way the measures can be reliable. On the other hand, what do you do?

You go back to CMS and say, look you have a whole suite of claims based measures, OK? And we know that administrative data is not very accurate, OK? Yet these things are being used to drive performance driven -- perform -- pay for performance, at huge impact.

And yet, the data elements really are not reliable. So, we want you to go back and show us that the data elements are reliable and they're going to try to do that, they're not going to be able to do that.

You know, then we're developing a whole suite now of measures that are based on the EMR. And how reliable is that data.

So, it's almost like we're starting from square one with, I mean, if we start telling people that they have to give us information about data element reliability, they could really throw a big monkey wrench in the whole performance measurement enterprise, just as a thought.

But getting back to this whitepaper, so, I think that what we're doing today and what we still need to work on is developing consensus on how to do this, and how to do it the right way, and I think it's really complicated and it's really hard.

And my guess is we could probably spend a weed doing this and we'd still have more work to be done. But at some point, we need to come to a consensus. And when we do that, I think there would be some value in putting that own on paper, OK?

And I also think that there would be some value in putting that through the peer review process and having those two pieces, one, consensus from this committee and, two, putting it through the peer review process I think would make it a better process for evaluating measures.

And once we have that process, then, again having it out there in the peer reviewed literature, I think would set some benchmarks for measure developers. I think it'd be important. I think it'd be a really important work product for our committee to do.

Male: I want to really support that. I think that for every reason that you said, it's also I think important – as I mentioned, I teach this stuff. I've never found any good material on this that I can use and I don't think it exists.

And I think that through training the next generation of measure developers, health services researchers and so on, it would be very useful.

Plus, it would be the beginning of a conversation with people. I mean, they may not agree with us, but then, they'll tell us why and so on. So, I think it could be very helpful in that purpose.

I also like to suggest the journal that published it in that eGEMs, the online journal that Academy Health publishes, came out of – I'm on the editorial board, came out of editor work at health centers). Maybe I do have a conflict of interest, right?

But it came out of the work that we did with (inaudible) because when I think about methods to use health data, not so much the substantive results, but the methods and I think this fits very nicely in that area.

Male: So, I completely agree with that. I guess to be more specific, maybe one of the things initially we could do is to conduct a systematic literature review to see some of the things like what have been done so far in, say, defining reliability and validity because those are the key terms.

And I think we can be very specific as to how we could proceed and then probably the next step would be to show it with public with data like how – so, if you go for this definition versus that definition like you know how we might get different results. So, I think I would very much on board with that idea.

Male: Can I – so, tried to catch him before he walked out, but it seems like we should have the writing – I mean I haven't heard any – this has come up before, the idea of a whitepaper or writing something and sending it for peer review possibly.

So, I think it's something that the committee wants to do and so should do. And usually the first step towards that is a writing committee because a committee of 20 people won't do a good job. But a writing committee of five I think can do a good job.

So, let's form a committee and maybe allow us – right here, but allow people to nominate themselves maybe to – Karen, or through the methods panel email that I think is mostly May. OK.

So, if you're willing to serve on the writing committee, submit your name. And if it ends up being 15 of us, then, we'll figure out, just to get a committee together to get something started, outlining, text, make assignments. We'll get the ball rolling, plenty to do there.

Female: And just to be clear, I think this idea of the toolkit and we're definitely going to get word to paper the idea. Peer-reviewed was a little new for me. We as an organization don't have many things that go out for peer-review. We probably should have more.

Male: And you may have to look into – like, you can self-publish, right? You put things on your website. So, maybe we should look into that. You put documents on you website.

Female: Yes.

Male: You self publish. So…

Female: So, we would definitely do that.

Male: You could possibly also arrange for peer review of your public. Do you know what I mean? It doesn't have to go into the journals that Mike suggested or anywhere else. But, maybe get the peer review but do it as one of your publications.

Male: So, this question came up internally at my place as to whether some of the things I put out were peer reviewed. I said absolutely. We peer review the entire country. If you want more people than any journal can get together to peer review this, it wouldn't be hard to do. You just – even outside of our own membership.

Female: OK. This idea of the toolkit, some of the things that we thought we definitely need in there are definitions, descriptions of methods, so, this would be pretty much trying to catalogue everything that might work and what it would work for and what the pros and cons are, those sorts of things.

Guidance on the best methods, so, there might be five different things that you could do, but this one is probably the best one and then, after that, if you can't do that, kind of work our way down.

Of course, this whole idea of thresholds, if we can get there or accept the real results or (rules) of thumb, whatever we can come up with would go in there as well. But are there other things that we need to be thinking about? Is that everything or is there more?

Male: Before we go into other things, are these the same now? Is the article – isn't that the toolkit or is the toolkit larger and the article is a summary of that or a subset of that?

Female: I mean, I think that would be one of the first things we'd have to come up with because a peer reviewed article could start out very, very basically and just talk about conceptually what is reliability and validity, and who you are.

I mean, you could do it very kind of high level and then eventually, you could have maybe other peer review that goes in some more detail or how we want to do it.

In my mind, the toolkit is this bighting that isn't static, that as we learn more and new methods come along, we add to it, whereas the peer review would be kind of here's where we are right now today whenever today is. They might need to be six months from now or a year, whatever you guys think. Is that…

Male: I'm just thinking about – so, the writing committee of the article, I think just so that we don't end up having a bifurcating process should be the same as the people working on the toolkit. And so, I mean – so, the toolkit is going to need a writing committee and I would think we would start them in the same place.

It may be that at some point in time, we say "this is suitable to send out for peer review and for publication," but it would be this sort of interim progress of the toolkit. It can continue on. OK. Otherwise, that could go in different directions and we end up saying, "How do we reconcile them" and create a new problem. Yes.

Male: Exactly that which is I see the paper being the toolkit at a point in time and then, the toolkit can continue to grow, change, et cetera.

Male: Karen? You're up.

Karen Johnson: Is that me Karen?

Male: Yes. I'm sorry. Yes.

Karen Johnson: I would actually argue for perhaps three products. I think the idea of putting out a perspective piece or a viewpoint, or whatever the particular journal might call it, it's a relatively brief sort of position statement saying "We as members of this new panel" plus NQF staff, whatever you're allowed to do in this realm "really want to make a statement about why it's important that all this stuff is happening, that this panel exists, that we're trying to get more systematic" to try to put out some sort of, I don't know, not technical document, but almost like a philosophical document saying, "Here's sort of what we've done in our first year" or two years or whatever feels appropriate to make such a statement.

And here's why it's important because I really do think that we're – as we sort of continue down this different era of quality measurement, quality measurement 2.0 or 6.0 or wherever we are that there is some import in that kind of a statement about the importance of methods and data science and methods evolution that is getting lost sometimes behind a lot of hype around data. So, I would advocate for sort of a perspective piece.

Then, I agree with the peer reviewed paper, but I would then think of the toolkit as actually something that's maybe a bit more staff letter, at least staff curated to be something that is set up to be more interactive and helpful to the user as opposed to as much of a narrative of "Here's some ways one could think about the different issues raised." So, to me, I guess it's three separate obviously need to be consistent products.

Male: I think that whole three-part thing makes a lot of sense. One thing the toolkit could include is lots of examples that developers could use as models. And

the peer review paper could refer to them, but obviously couldn't have enough detail to be as useful as the toolkit would be.

Female: A very obvious link but something I have not thought of. So, thank you.

Female: Yes. Just a clarification, so, you obviously have a lot of pieces of this toolkit. Do we intend this to sort of leverage what you have, what you've already provided to us, start from stretch? I mean, how are you seeing this, something new? What are we thinking here?

Female: I mean, I think we definitely have our definitions and I don't think we really changed our definitions tremendously based on today's discussion. So, I think we have our definitions. In terms of what NQF has put out in like descriptions of methods and things that are appropriate, we actually don't have anything written.

We have kind of our collective experience on the types of things that have come in and things that we've kind of let through or haven't let through and all that kind of thing. But we've never actually had – we've always had the will to do it and we've never really had a vehicle to make that happen.

So, I think it's a lot from scratch or a lot drawn from pieces here and there. And in terms of the guidance and thresholds, we don't have those at all.

(Paul Kurlansky): Getting some thresholds is something that tortures me a little bit with this process. Even if you get a measure reliability and it's like 0.5, so, that's – what do you do with that?

I'm a little more comfortable in the world of risk models and whatnot and C statistics and things like that. But at what point do you say that the reliability is acceptable? And I don't have an answer.

Female: Lacy?

Lacy Fabian: Yes. Just building on what Paul said, I think fundamentally the theory of having the threshold is something that seems incredibly long overdue and would be really helpful.

But I feel like that undertaking would almost just be a whole separate in and of itself. Like, it's a vision. I mean, if it's going to take us two weeks to figure out the data elements, I think threshold…

Male: Well, I think – I've had a little experience with this getting pressured from sometimes the community, other times, the journal editors saying "Give us a threshold for this event."

And we resisted. But, we've come up with a language where you can put a number down and qualify it with enough nuance that you're not saying "0.4. If it's 0.39, it's out. If it's 0.41, it's in." But if it's a general – and I'm not suggesting 0.4 by the way, I'm just using it as an example.

But, if there's a general in that range or maybe you could say, well, 0.35 to 0.5 or something, then, that's kind of a range where depending on other aspects of context and application and sample size and in fact it's -- and I think we could pull that off as we get into the weeds on that.

At least I think there would be a – I hope that we would and developers at least then have some sense of how well it's going to fare if they send it in and maybe how they need to explain it more vigorously if it's on a low end.

Male: Yes. A question for the NQF staff, what level of engagement would your board need? And I'm thinking more around the idea of like threshold, that's coming out with recommendations for thresholds.

I mean, to me, that has significant implications to measures and what measures could even brought forward. And so, I just am wondering if we would need to engage them early on that conversation.

Female: Yes. It's a good question. We did start engaging the consensus standards approval committee. That's the (CSAC) that oversees the CDPs, endorsement committees. In the past, it would have been the board making this decision.

But since that decision has been brought forward to the (CSAC), we've been trying to socialize, "What do you guys think if we start having thresholds slowly?" This is the way to kind of get them on board.

So, we do have a meeting with them in about two or three weeks and we'll have a summation of this discussion with them so that they can get comfortable with that. But, as an organization, what we're trying to do is make sure that we are prioritizing the measures that really matter and this is one way in which we're doing that.

We do have criteria that speaks to that, but it's also through the assessment of the measures and the evaluation of the measures. So, it's very much in line with discussions we've have with them recently.

Male: David? David Behrens is on the phone again. So, welcome back, David.

David Behrens: Yes, I am.

Male: Thank you. And Joe has been waiting on the phone to raise his hand (and, Susan). So, go ahead, Joe.

(Joseph Kunisch): Yes. As I was listening to everything, I was thinking and I had suggested this in another panel meeting, would it be feasible to come up with some kind of scoring or weighting measure for a quality measure.

So, I'm thinking about like right now there's a funding request out for the quality payment program developing quality measures. And if you go into the actual application, each section has the point value assigned to it.

So, every section that you can fill out, there's 10 points or 20 points. So, you can kind of see where CMS is weighting more like does it fulfill a gap that they're looking at and so forth.

So, I don't know within this toolkit if that would be something feasible that would give guidance to a measure developer to know how strong their measure was when they were going to submit that application to say, "I hit 90 points out of a possible 100" and also a way to give them feedback to say, "Well, this section, reliability only scored 5 out of 10 points." So, you know if you're going to resubmit this, this is where you should focus.

Male: I think that's a good aspiration for us to put into our hope list.

Female: We can put that – that sounds harder than thresholds to me.

Male: Yes.

Female: It may not be, but…

Male: We can strive it and keep it in front of mind.

Go ahead, Susan.

(Susan): The only suggestion I was going to make on thresholds is sort of the idea of treating it like how good of an ACT score do I have to have to get into Ohio State.

We say generally we accept people at X number and above and we have enough measures certainly that have been accepted that we can say "A C statistic of this is generally acceptable, but it's on balance with other – with your GPA and your sports and your activities," right?

And so, I think if we treat it – I would hate to treat a threshold as a hard stop. So, I think if there's a way we can look at this now that you have enough history, if we can look at that and say, "Hey, things in this general range go through."

Male: Thank you. That's kind of what I was trying to get at, rather than come up with – I think if we try to come up with an absolute number, we'd never converge on a number. And then, once we did, we wouldn't feel good about it. But, it is also useful to provide contact information and…

Female: On what kind of metadata we have around the applications. Yes. So, that might be something worth considering.

Male: Yes. I mean, we probably could generate what's the average correlation coefficient of ones that fail with reliability and the average of the ones that succeed.

Female: Yes, the (ICCs) or whatever statistic they're using, we kind of…

Male: Yes.

Female: We can start simple. I'm not saying go through and extract 600 statistics on every app. I mean, there's probably three or four things that you can grab that might be informative.

Male: Yes.

Female: I know when I was in the developer side for a little – I tinker over there, like most of us do, and if you get a measure and you see an ICC that's like – I don't know, that doesn't look that great, but then you find out a bunch of other measures were worse and they made it in. You don't feel, right? You're like, wow, I'm not even close. I thought I was great.

Male: There's hope.

Female: So, I just think it'd be helpful for the submitters maybe.

Male: Yes. Yes.

Female: And just a couple of things to your point, you may have noticed, I think every place on the slide deck where I have the word "threshold", I have it in the quotes because of that exact idea that probably having a hard number may not be even possible or maybe it's more possible for certain things than others, I mean, I think we'll have to go through.

We were giggling just a little bit. We have kind of started trying to figure out because we could look back and find the experience. It's not quite as easy as we would like it would be. So, it will be a slog through the paper. Yes.

Female: The form has been a bit of a work in progress.

Female: Yes. Yes.

Female: So, you might have to cherry pick things that have stayed fairly consistent.

Female: Yes.

Female: I was just going to call it out directly to the point about the drawing out the data. That can also inform our consensus side of things, too, if we've got people saying "This is fantastic and great" and the statistics was not, and others who were like oh, those sort of folks.

Female: Yes. I mean, a part of what is interesting at least in our first two cycles, our prediction has I think come out which is we expected you guys to be harder graders. So, we'll have to remember that as we look back on what went through before. We wouldn't want to say that just because it went through before, it's reasonable to go through now.

Male: I just had one thing. Actually, I see bullet points three and four ss being really helpful for us which is the guidance on best methods for different measure types and the thresholds are acceptable results.

I see that being really helpful for our group as we work towards hopefully higher (inter-rater) reliability if we're all sort of clear on these are the methods we should be looking for and these are the ranges we should be looking for.

Male: You'd be the first users of the toolkit.

Female: I was just going to add that I think the data – the history data will be interesting. But I don't think we should read too much into it because that's the reason – one of the reasons we put this committee together was that you said a lot of decisions that were made in the past were made by people who didn't necessarily have the skills or the knowledge to make those decisions and they might not know what ICC point two meant.

So, we might see – I would expect we'll see a whole lot of inconsistency there and a lot of decisions that we probably wouldn't have made.

Male: So, I mean, just bringing back this comment about claims data and I mean, it's a little bit like the statement that if aspirin went to the FDA today, it would never get through and we would never have aspirin. If the claims data measures went through today, they wouldn't make it through those committees.

So, what is the – I mean, how are you going to deal with that? Because if it goes forward in three to five years, you're going to have new measures that meet a higher standard and then a lot of legacy measures that don't. Maybe that's a question, not a fair question, but…

Male: There is always the maintenance process. Everything comes back for review within three years. But I think we'd set it up so that it doesn't always come to the methods panel if it's all been previously approved, but maybe we shouldn't rethink that.

Female: Yes. What we said is it won't necessarily come to you but unless it has new testing. And most of our outcome measures luckily do. We usually have new (processes), so more than likely that you will get those things.

Male: More of a philosophical answer, but my answer to that would be things get better. In other words, it's an imperfect world. It's an imperfect science.

I look at things that I published 10 years ago, I'd be ashamed to submit them today. It's a different world. But I'm just saying. So, yes, there will be legacy metrics that linger and probably aren't ideal but we should just try to make sure that in the future, we get better.

Male: Well, I mean, good point. One difference though, once it's published, it doesn't go away. But some of these legacy measures can go away and maybe that's just something to think about three to five years from now.

But, if – and you can either sundown because there's no longer variation that's worth measuring and grading on or because the measure is not measuring up. Lacy?

Lacy Fabian: I think just building on your point, maybe that comes back to this toolkit and even this peer reviewed journal paper is putting out into the measure developer community and trying to be very transparent that these changes are afoot. We are expecting progress.

And being transparent about what those new criteria are and how things are going to be expected in the future can set the stage or raising the bar in

making those shifts in those legacy measures, that need to have shifts than where it's warranted or sunsetting them if that's the case. They don't all have to come back for maintenance if that's not the case.

Female: I don't see anybody else having -- anybody on the phone want to say anything?

Dave Behrens: Just quickly, Dave Behrens here, I appreciate your patience with me and these different places. Basically, I agree with the line of discussion we've had on a number of points.

And the idea is talking about ranges of things that have been accepted or rejected in the past with as much new bussing up as we've given them as opposed to finite thresholds which I think was good.

It is kind of a problem that there maybe is something that's gone through in the past that we would think are quite acceptable now.

But there's probably ways we could deal with it because the messaging could just say that even though this has been – we're describing characteristics of measures in the past, this doesn't necessarily mean you will see this all going through and maybe then we have to qualify as things are becoming more stringent.

And a general point is that I think we can talk about ranges and give people general guidance without throwing out specific target numbers that tend to be then accepted as really black and white. So, I like the direction of the discussion. I think we're on the right track.

Male: Well, we're kind of reaching a closing point. The next steps, I mean, maybe we should – there's a public comment coming. I don't know if we need to wait to set that time or allow that sooner. But before we do that, should we get a little more specific about next steps on the toolkit? What do you think?

Female: Before we go there, did anybody have anything that didn't make it to the parking lot that you want to make sure it does because the parking lot is also going to be our next steps, too.

So, we kind of have lots of different next steps, right? And some of the things, for example, that we have to plan is what are we going to talk about on our monthly call.

So, that is even potentially a next step. So, did anybody forget to mention something that we want to get on the board? Yes. If you want to read that out, I know I have a few things written down that I may be had a pin on.

Female: So, things that we have on the parking lot, first is the issue of Q.I. versus accountability, treating those measures differently potentially and so on across between what developers submit and how we review those measures, data elements for multi-items (of scales), individual items versus rolled up items and kind of the confusion with that, and a measure score versus data element, VRK of data element and measure score, so, that question of did they appropriate data element? Are they using them the way they say they are? And then, are they combining them properly?

So, kind of a different level of questioning that you want for reliability – is precision really equal to reliability? Does it not equal reliability kind of quite the measure score? And is signal-to-noise enough of an indication of reliability?

And then, reliability in the context of use and the last one was the lack of variation among entities does not always equal a lack of reliability, going back to that I think David's earlier comment about, well, everyone is doing badly or everyone is doing well, that lack of variation doesn't mean that the measure is not liable. It just means it's reliably capturing that everybody is equal.

So, those are the things that we have in the parking lot, any items that we've forgotten. And this did not include any of your suggestions for process improvements and such.

Male: Well, I don't know if this is the best way to do this, but it is a way and what we could probably learn pretty quickly if it's not the best way and that is to get a glossary.

I mean, just get the piece of paper in front of us, write the term down, and put the NQF definition of that term whether it's precision or stability or all the other terms that came in reliability and validity. Just, how does NQF define them?

I know we have some of it and we have tweaked it and/or accepted it. But, let's – that's a parking lot item, right? I mean, it kind of covers some of these things as to – and it might be that the process of doing that and kind of vetting it with the committee will help shape the outline of the toolkit because it may very well be that the way to really frame the toolkit isn't around the glossary but around the measure and that could help the writing committee of the toolkit.

And I don't know if a writing committee is a parking lot item or just a different action item. But I agree but we're going to get a writing committee together.

And I guess my inclination would be to start with one group and if that group says, "You know what? We really just want to work out on the article. We rather have somebody else working on the toolkit and we'll sync up."

Then, maybe we need another committee on that. But I would suggest we start with one group that get going and then work with the larger group and just keep reporting on where they are and where they're heading.

And if they're leaning more towards we're really developing the toolkit because we think that's where we should start. The article will come from that. We really want to focus on the article.

But, it's one way to do it, unreasonable? So, form writing committee, that's a parking lot and whether they're writing the toolkit or the article or both or there was a third – the overview piece, the positioning paper.

Male: Which could potentially be a different group, right, because that feels…

Male: I think that group on a position paper could be – I would think that needs to be kind of heavily represented by NQF because it's going to be like your – it's our

position but it's our position as a panel that you deploy. Yes. So, maybe we could…

Female: Figure out.

Male: Yes. Yes. Yes. And you already have text on that on the website and – yes.

Female: I think all of ideas are helpful and I think (inaudible) we need to go back as a team to kind of think through it, which ones can we take on and how quickly and timeline wise and then really what is the goal each item and then who really is involved.

So, I think we're going to need to think thoroughly about it and figure out the best process and obviously find a way to keep everyone as involved as possible without obviously having everyone to write it.

Male: So, the message I heard was you have other aspects of your job. This isn't your entire job. You're just serving this committee.

Female: Yes. I feel like we've made a lot of progress. We covered what we wanted to cover pretty much. And I'm happy with where the day has landed. I don't want to drag it out too much more.

Does anybody have any suggestions you want us to think about as we kind of go forward? We've already talked about processes, not that so much, but are we pretty good? Yes.

Male: Should we maybe open up the lines for public comment, please?

Operator: And if you have a comment at this time, please press star, one on your telephone keypad.

Female: One more question for you.

Female: Yes.

Male: Does anyone would like to comment?

Operator: There are no comments.

Male: Thank you.

Female: OK. When we asked you guys to help us with this, we said we're going to have an annual meeting in D.C. and calls every month and it's going to be a one-hour call. Do you guys feel like our one hour call is long enough? So, is there any appetite for making it a two-hour call? And it's fine if you say no.

Male: I'd vote for one hour.

Female: OK.

Male: I think it sort of forces people to be focused and I think we will get as much accomplished in one hour than two hours.

Male: I can go for the middle ground, one and a half hours.

(David Behrens): Yes, David here. I think part of our discussion earlier today brought up the possibility of two in-person meetings rather than one and I know that's not a settle point.

But if we do go in that direction, (inaudible) that the public calls are a little more – try to be shorter (inaudible) and could be kept to an hour. Who knows? (Inaudible) every single month, but there seem to be some support to the idea that a well structured (inaudible) out to where it is.

Female: OK. That's what we needed to know. Next steps, is that next?

Female: So, the (inaudible) spent today taking copious notes. And as has been alluded to, we will regroup as a team and figure out how to best operationalize all the feedback and recommendations that came out of this meeting.

The methods panel will reconvene next on June 14th for a one-hour monthly call but in the interim, please feel free to reach out to the methods panel inbox if you have any questions or concerns. And I'll turn it back over to the Chairs for closing remarks.

Male: Well, I think it was a great meeting. I think I've heard from people during the break and just sort of same thing with faces around the room that we're making progress and I think there's little considerably more – more than minimally clinically important magnitude, more comfort with where we're going than was evident in the first time we met when we were all still figuring out who we were and what we were doing and what these terms meant.

And I'm looking forward to seeing things on paper and then having a discussion around that because I think that will elevate the discussion more to a level of what can you live with, not just what you have to say, and what do you think, but what can you live with.

And then that will open up to a wider group when we send it out for public comment or for peer review. I look forward to that process. Karen, do you want to make – yes, David, if you're still, if you haven't boarded the plane yet, do you like to make any closing remarks?

(David Behrens): Thanks. I kind of hear the overhead announcement. We are boarding (inaudible). I thank everything, great meeting, good focus, a lot of comments. I like the (inaudible) So, I'm very encouraged every day and I just thank everybody for being here and your contributions today.

Male: Thanks, David.

And because Karen Joynt Maddox has been our able leader for up until this meeting, we want to give you a chance to make any closing comments as well, Karen.

Karen Joynt Maddox: Well, thank you. No. I just really appreciate being able to remain involved. This has been a really neat day.

I took a lot of notes and learned a lot and I really feel like this is going to change the way that we think through things and hopefully help guide the way that measure developers and quality programs sort of collectively help move this field forward. So, I'm really excited about it.

Female: So, I get to say the last word and my last word is thank you so much for coming, for being such giving people. This is a lot of time on your side and we really do recognize that and we very much appreciate it. We're very grateful.

And I personally am a bit of a geek, a methods geek. So, I've had a blast today. So, hopefully you guys have as well. I'm exhausted but I have had a blast. So, thank you guys. Karen, best of luck, we probably won't be hearing from Karen on the 14th. Maybe, I mean, I don't know, but good luck with the birth and safe travels.

END