The Expert Panel met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Elizabeth DeLong and Patrick Romano, Co-Chairs, presiding.

PRESENT:

ELIZABETH R. DELONG, PhD, Duke University Medical Center
PATRICK ROMANO, MD, MPH, UC Davis School of Medicine
JOHN D. BIRKMEYER, MD, University of Michigan
DALE BRATZLER, DO, MPH, Oklahoma University Health Services Center*
JAMES CHASE, DO, MPH, Minnesota Community Measurement
NANCY DUNTON, PhD, FAAN, University of Kansas Medical Center, School of Nursing
ELIZABETH GOLDSTEIN, PhD, Centers for Medicare and Medicaid Services
SHERRIE KAPLAN, PhD, MPH, The University of California - Irvine
LYN PAGET, MPH, Health Policy Partners
DAVID SHAHIAN, MD, Massachusetts General Hospital
STEVEN WRIGHT, PhD, Veterans Health Administration

ALAN ZASLAVSKY, PhD, Harvard Medical School

NQF STAFF:

HELEN BURSTIN, MD, MPH

KAREN JOHNSON, MS

ELISA MUNTHALI, MPH

* Participating by teleconference
Welcome

Eliza Munthali...................... 4
Elizabeth DeLong................... 5
Helen Burstin....................... 5
Patrick Romano...................... 6

Introductions and Disclosure of Interests

Helen Burstin....................... 5

Overview and Context

NQF Staff............................. 16

NQF Experience with Composite Performance Measures

NQF Staff............................. 28

Guided Panel Discussion

Elizabeth DeLong and
Patrick Romano....................... 73

NQF Member and Public Comment...... 215

Applying NQF Measure Evaluation Criteria to Composite Performance Measures

Measure Submission Form, Review and Evaluation.............................. 220

Measure Submission Form, Review and Evaluation.............................. 362

NQF Member and Public Comment...... 374

Next Steps............................. 375

Adjourn................................. 378
(8:34 a.m.)

MS. MUNTHALI: Hello and good morning everyone. Welcome to the Composite Measure Evaluation Guidance in-person meeting.

My name is Elisa Munthali and I am a Senior Project Manager with NQF.

Before I turn the meeting over to the technical panels' co-chairs, there are a couple of housekeeping items I just wanted to bring to everybody's attention.

I wanted to remind everyone that today's meeting is being recorded and transcribed. And so we ask that when you are speaking, you turn on your microphone and speak into the mike, so that we can capture your comments.

And for everyone that is in the room today, your received a packet of handouts. I ask that you keep those handy. We will be referring to those and we will let you know which documents we are speaking
about. And also for everyone in the room, we
just wanted to let you know where the
restrooms are. They are just beyond the
elevator right by the reception area there and
through the glass doors. And also just to
remind you that we have breakfast in the back.
So please help yourself throughout the morning
today.

So I will turn it over to Helen
Burstin who will continue with our welcome and
conduct our disclosures of interests.

DR. BURSTIN: Great. Hi,
everybody. Helen Burstin. I think I know
everybody. I am the Senior VP for Performance
Measures at NQF. We will go around and do
introductions and disclosures at the same time
-- there she is. Excellent planning.

Before I do that, why don't I ask
Liz and Patrick to introduce themselves?

DR. DE LONG: I'm Liz DeLong. I'm
the Department Chair of the Department of
Biostatistics and Bioinformatics at Duke
University. And I have been at Duke for years and years and I have worked with Dave Shahian and Patrick but I have heard the rest of your names and I think you are probably all much better at composite measures than I am at this point but we will see.

DR. ROMANO: Hi. And I am Patrick Romano. I know most of you. I am a general internist and oral pediatrician based at UC Davis Health System in Sacramento, California and long-time health services researcher involved in Quality Measurement. I worked with NQF before on several previous projects, as well as with AHRQ and other organizations in the field. We will talk a bit more, I guess we will have disclosures again and I will make separate disclosures.

Anyway, it is a pleasure to be here and thanks to NQF for convening us for this purpose.

DR. BURSTIN: Great. Well thanks again. So what I would like you to do is as
we go around the room if you could introduce yourselves. You may remember you filled out a disclosure of interest form which you sent to us. You don't need to fully go through all of that. I think the key thing at this point, since we are not really evaluating measures today so you can't have any conflicts with specifics measures, is really just I think more than anything else to give a sense of sort of where you are coming from for your fellow committee members. And if there is any areas that you think are important for people to understand in terms of potential bias, everybody has got opinions, obviously, we all are here today because we have opinions and do research and feel strongly about things.

So as you are going around the room, introduce yourself, where you are from. If you think there is anything relevant that your co-committee members would want to hear about, feel free to mention that. I also welcome Karen Johnson, our Senior Director who
made it off a very late train from Maryland.

So thanks.

Start with you, Steve?

DR. WRIGHT: Hi, everyone. I'm Steve Wright. I'm with the VA Department of Veterans Affairs, Director of Epidemiology currently acting as the Director of the Office of Performance Measurement. I have been involved in measurement for many years, health services researcher as a background.

I don't have any particular biases, other than rah, rah, VA!

(Laughter.)

DR. ZASLAVSKY: I'm Alan Zaslavsky. I am a Professor of Health Care Policy Statistics, there is a parenthesis there, at Harvard Medical School, Department of Healthcare Policy. As my title suggests, I'm a statistician. I have done a lot of work particularly on the CAHPS survey since the inception of that project and also with the HEDIS measures.
DR. CHASE: Good morning. My name is Jim Chase. I am President with Minnesota Community Measurement. We are an organization that does quality measurement around the state of Minnesota and have about 600 medical sites of care that report data to us just about every provider in Minnesota.

And we use composite measures. We have a couple that are endorsed by NQF so that would be my probably major bias and this is just the experience in using the all-or-none composites in our community.

MS. PAGET: Good morning, my name is Lyn Paget. I'm down from Boston where I have recently started some work with a group called Health Policy Partners and Independent Collaboration of Patient Policy Experts. I spent many years at the Informed Medical Decisions Foundation. So I am here much less as a measurement expert but more so in the position representing the patients and consumers whom we hope will benefit from these
measures.

DR. DUNTON: Good morning. I'm Nancy Dunton. I'm a research professor at the School of Nursing at the University of Kansas Medical Center where I direct the National Database on Nursing Quality Indicators. We collect data quarterly from 1900 hospitals across the U.S. on structure, process, and outcome measures relating to nursing care and we think about composites.

DR. GOLDSTEIN: Liz Goldstein. I'm Director of the Division of Consumer Assessment and Plan Performance at CMS. I have been involved in the CAHPS surveys for many, many, many years. My division is responsible for most of the CAHPS surveys that CMS implements as well as we are starting to develop three new patient experience surveys. They are just starting up.

My division is also responsible for the star rating system for Medicare Advantage. So we use lots of different
measures for that, as well as we create our
own composites for value-based purchasing for
Medicare Advantage.

DR. KAPLAN: I'm Sherrie Kaplan. I'm a psychometrician by training. I'm Assistant Vice Chancellor for Healthcare Measurement and Evaluation at UC Irvine. And I have been working in creating composite measures most distantly at Rand with a medical outcomes study. Most recently with the State of California, I am trying to help advise them how to sample, whether to sample more items per constructs, more patients per doctor, more doctors per clinic, more clinics per institution, et cetera, et cetera, and making the most out of how we look at these composites to do institutional performance assessment.

DR. BIRKMEYER: Good morning. John Birkmeyer, I am a researcher from the University of Michigan. I direct the Center for Healthcare Outcomes and Policy there and
as a researcher I have been engaged for many years with colleagues, economists, Doug Staiger at Dartmouth and Justin Dimick from University of Michigan with more statistically-based types of composite measures looking mainly but not totally at surgical care.

I have been involved with the Leapfrog Group for many years and have been involved with implementations of its composite measures. Most recently with Patrick, the hospital-wide summary composite score for patient safety and earlier for more statistically-based composite measures for surgical standards.

By way of disclosure, I am the founder and Chief Scientific Officer of a company called ArborMetrix that is not a developer of measures but we do implement performance measurement systems for insurers, health systems and for professional organizations.
DR. SHAHIAN: Hi, I'm Dave Shahian. I'm at Mass General and Harvard Medical School. I chair the STS National Database and its Quality Measurement Task Force. I have been involved in development of two cardiac surgery composites, one of which is NQF-endorsed and publicly reported for CABG and another one for isolated AVR, which is just, it is going to be published in the peer review literature next month and will also be publicly reported and submitted to NQF.

I have no disclosures. My positions with STS are uncompensated.

DR. ROMANO: Okay and just to amplify a little bit, obviously I am an employee of UC Davis Health System but I also have done fairly extensive work as a subcontractor to Battelle Memorial Institute working on the AHRQ Quality Indicators program. So in that capacity, I have been involved in some development testing and application of the AHRQ QI composites, three
of which are currently NQF-endorsed. I have also, as John mentioned, been a member of Leap Frog's expert panel related to its hospital safety score composite program. I have also advised the California Office of the Patient Advocate, which is responsible for health plan and medical group reporting in California related to construction and reporting of measures, including a compositing of measures.

And I have also done a little bit of work on expert panels for a variety of organizations, including CMS, as well as the AHRQ and probably others -- Joint Commission and probably others I'm not thinking of.

In any case, Liz?

Dr. De Long: I worked with Dave on the composite measure for the STS CABG surgery and I have primarily focused on outcomes research and have no ties with industry or other developers.

Dr. Burstin: Dale are you still with us on the telephone? Can you introduce

---

Neal R. Gross & Co., Inc.
202-234-4433
DR. BRATZLER: I am. I am, thank you.

DR. BURSTIN: Yup.

DR. BRATZLER: My name is Dale Bratzler. I am a professor in the Department of Health Administration and Policy in the College of Public Health at the University of Oklahoma and also a Professor in the College of Medicine.

I have worked on the development of performance measures for many years, primarily as a contractor to the Medicare program. Currently I do have contracts that are through the College of Public Health to support continued measure maintenance for the Medicare Program. I also have a contract to support external quality review activities for the State Medicaid program.

My work in composite measures has really been relatively limited. We have developed composite measures that we have used...
over the years, primarily to support performance improvement. We really have not developed composite measures that we have ever recommended for formal endorsement or public accountability. So we have been using composites for some time but primarily we have been using them to help drive performance improvement.

DR. BURSTIN: Thanks, Dale. Does anybody have any concerns about anything they have heard about each other? Any further questions or probing or are we ready to go to work?

All right, great! Thank you everybody. I guess at this point, you are five minutes ahead of schedule, Karen. This is Karen Johnson, who is our Senior Director on this project. Karen Pace, who many of you was hoping to be with us today but unfortunately has a personal urgent issue and can't be with us. But Karen is well primed and ready to go.
MS. JOHNSON: So thank you. And I am sorry I am running late today so I haven't got to meet you guys but I am really honored to get to work with you. Composites are new to me so I get to learn from some of the best. So I am really excited about this.

So we wanted to give you just a little bit of overview and context about NQF's experience with composite measures. And just to remind you, in 2008 and 2009 NQF convened a TEP to identify a framework for evaluating composite measures and during that project, a definition was created or developed and principles were articulated in terms of how to evaluate composite measures. And also specific criteria were developed so that we could evaluate measures as they came in. And at that time four AHRQ measures were evaluated and kind of served as a dry run for our framework and evaluation criteria.

So since that time, however, we have updated both our criteria and our
guidance for our other criteria that we use to evaluate measures, specifically our evidence and our scientific acceptability, reliability validity guidance has been updated.

So what we need to do with this project is pretty much re-think our guidance for composite measure and evaluation and make sure that it fits with our updated guidance that we have for our regular measures.

So with that in mind our goal is to update our guidance. So the three things that we would like to accomplish with today's meeting or to identify appropriate evaluation methods for various types of composite measures, identify unique considerations for evaluating composite performance measures in relation to our endorsement criteria and then finally develop guidance for evaluating and submitting composite measures for endorsement.

So the actual nuts and bolts of what our submission forms look like.

So you have some resource that we
have tried to provide for you. And Elisa
thank you for getting all this printed out in
my absence this morning. Beyond your
expertise, which I know you have bucket loads
of, we have provided you an agenda so you know
where we are going today. The briefing memo
that Karen wrote, which I think we will
probably follow that a lot as we go through
our agenda for our meeting.

We have provided you our measure
evaluation criteria so that you know what we
are trying to align with. We have also given
you our composite criteria so you know what we
currently have now and we have also given you
kind of a not very pretty but I think it
covers the basics of our composite submission
form so that you can see the actual questions
that we asked developers to fill out when they
submit a composite measure.

Just a little bit more context
beyond what I have already mentioned. All
NQF-endorsed measures are considered suitable
for both performance improvement and for
accountability. So currently we do not
endorse performance measures for specific
accountability applications. And then also
the term composite measure and even the term
composite means many things to many people but
it can refer to scales, or instruments to
assess individuals or performance measures
used to assess providers. So just a reminder
that we endorse the performance measures, not
the instruments or scales.

Okay, so on to our experience.

Yes?

DR. DE LONG: Can you go back?
I'm not sure I understand what you mean by NQF
does not currently endorse performance
measures for specific accountability
applications.

DR. BURSTIN: So essentially when
a measure is endorsed by NQF, at least at this
point in time, the assumption is it is ready
to go for any purpose. If somebody picks it
up for payment, if somebody picks it up for public reporting, that is the assumption out the gate. What we don't do is distinguish this measure is great for QI but it is not quite ready for payment. And it has been an issue that keeps coming up in a big way, as Sherrie knows well from our All-Cause Hospital-wide Readmission Project.

And it may be something that will morph over time. Currently there is another partnership called the Measures Application Partnership that actually helps to think through specifically which applications are appropriate for which measures and which federal programs. But it is an important distinction just because people often times say this measure is great for this purpose but I wouldn't use it for that. And at least at this point in time, that is not the way we can really kind of separate out that thinking.

DR. BIRKMEYER: So just to be clear though because this is so crucial and
more crucial to the composite measures than any other type of performance indicator, the measures need to meet some low bar that it could be used either for improvement purposes or for accountability or steerage but recognizing that obviously some measures are going to be much better suited for one or the other. Is that fair?

DR. BURSTIN: That's fair, yes.

DR. ROMANO: Well yes, I guess maybe you could elaborate on this a little bit more but the fact that there is a measure applications partnership that is trying to think systematically about the application of largely NQF-endorsed measures in various accountability applications implies that the existing NQF process doesn't really comprehensively consider the implications of different accountability or specific accountability applications.

DR. BURSTIN: Right. So there is an assumption that if it has been endorsed by
an NQF committee, endorsement side committee, 
that they are primarily focused on the 
measurement properties. They are really 
looking at the criteria that we use to assess 
the measure. They are not looking to say, for 
example, this measure would be more 
appropriate, to put it in real terms, this 
measure would be great for the hospital public 
reporting program but we don't think it is 
ready for value-based purchasing. That is not 
something we do as part of the endorsement 
process. That is currently something done as 
part of the MAP process. 

Again, I think we are increasingly 
-- it is an interesting time, as all of you 
know, in a big way. And if you look to the 
example of Massachusetts, for example, they 
did come up with some criteria of what our 
higher stakes measurement criteria might be. 
It is not something we have explored yet but 
it is certainly something, it is hard to deny 
a sort of back of mind of seeing for example,
the MAP last year put one composite as being okay for the hospital public -- you know for the IQR, hospital public reporting and yet said it was not okay for value-based purchasing. That implies a difference that people are sort of thinking through. I don't know that we have teased exactly what it is that the MAP is using to make that decision, other than the multi-stakeholders kind of talking it through. But I think it is an important issue and you are absolutely right, John because obviously composites are pretty high stakes. And you put it all together and you say this is safety or this is high quality for diabetes. It does have potentially a different lens. And they do tend to be picked up for, I would argue, sometimes higher stakes applications than I think some of the other more individual process measures might be.

MS. PAGET: So Helen, can I just ask then do you think it would be part of the role of this group today to make any
recommendations about that or do you think that that is in the MAP domain and we don't need to go there? Because I find myself looking at this and thinking about the influence of these measures just on cultural change and the care experience, which is really a completely -- and as I read through that one paper from the Research and Battelle, it was that one example of just using these to choose a physician, well that is just the tip of the iceberg in my mind. That is really probably not the way these are ultimately going to be used.

So for clarification, I just wanted to know whether you think is something we are going to discuss today or not.

DR. BURSTIN: You know you are really smart people sitting around a room. I think anything is fair game. At the end of the day, we want to make sure we accomplish the goals of saying what do we do when these measures come forward to us because we have
seen every stripe of composites come to us and at times we feel like we are really pounding a square peg into a round hole. So anything you guys could help us think through with clarity that we really do have an approach that makes sense. If you have other thoughts about these other issues, again, we would be open to hearing about it, as long as we get the rest of the work done because I honestly don't know what will come forward. I think there is just going to be a lot more movement in the next couple of years of just clearly seeing a sense of high-stakes measurement versus not in understanding how to handle those.

But at this point in time for where we sit at this point, we should assume that anything that is endorsed by NQF if appropriate for any of those applications.

MS. PAGET: It does work for me but I actually had another question on this side which came up with the notes on the
composite measures that have been submitted
and I think it was in reference to CAHPS where
the statement is made that NQF doesn't look at
the survey instrument itself. I just need to
understand.

    DR. BURSTIN: It's not that we
don't look at the survey instrument itself.
Obviously, Liz and Alan and others can help
here but essentially at the end of the day
what we are endorsing is the performance
measure based on use of the survey. So we are
not endorsing the CAHPS tool, per se. We are
endorsing a performance measure based on
CAHPS.

    So this is a big issue it says at
the bottom of that slide there "see PRO
project." Some of you may know we have been
doing extensive work over the last six months
or so on a project around patient-reported
outcomes and really trying to tease this out.

    So actually as part of that
project, we would be happy to show those
commission papers with you. It is a very clear distinction between the PRO, the patient-reported outcome as the tool, and then ultimately how do you use the tool in the context of moving to a performance measure using the tool and actually tried to come up with the critical path of how to get there.

MS. JOHNSON: Okay, so from 2007 until now we have had 28 measures that have been submitted to us that have been flagged as composites. And just to put that into a little bit of perspective, since 2010 we have evaluated more than 400 measures, so 28 out of 400, it is a small number but they are difficult because they are complex and that means that some of us working on the projects maybe have never seen a composite come through. So even getting clarity about these criteria is going to be important to us as staff as well as developers.

So of the 28 that have been submitted, 22 are still currently endorsed.
And you see the breakdowns there. We have three that are all-or-none measures, six based on CAHPS surveys, five on other types of surveys or instruments, and then the remaining eight are combinations of this, that, and the next.

So of the 22, some of those have been through endorsement maintenance but not all of them. And I think one of questions that Liz had was in terms of are they useful. At the end of the day, have they been able to improve quality? And we would love to be able to answer that question systematically. Right now we can't just from looking at our data. We do have institutional memory in terms of Helen and Heidi and different folks who sometimes can chime in on these kind of things. But right now, other than if you have some examples that you might want to mention, Helen.

DR. BURSTIN: Right. So two things. So the first is that we recently
updated one of our four criteria of unusability, which is now usability in use. And the idea there was to really much more carefully delineate what do we mean by when a measure is usable. And it is really implying that there is significant benefit in terms of driving quality improvement, improvement in performance results, but also not forgetting about potential negative consequences of measurement as being now woven into that very clearly.

So as part of that, and that is now coming forward for all measures being submitted, new in maintenance to NQF, a very clear requirement to say of your data how has it helped? Has it moved the needle? Has it potentially hurt as part of that process? But to date, much of what we have gotten from folks has been, here is the measure. It is used in four states. Here is the measure, 16 other groups are using it. We don't often get and here is how it moved the needle, actually
with the possible exception I would say, Jim, of some of what we have seen from Minnesota Community Measurement when the cardiovascular committee re-endorsed the optimal vascular care composite, Minnesota was able to give pretty strong data showing significant improvements in cardiovascular outcomes in a way that we don't tend to have. But they are above average in Minnesota.

But that is the kind of information we would love to be able to get more systematically. So as you are going through this review today, helping us think through in particular about the composite issues, and one of the things we have often heard a lot and it comes up from our consumer purchaser council in particular is if a composite is put forward and is publicly reported, is it also unpacked? So is it unpacked for those obviously trying to improve care in terms of being able to see the individual results but is it also unpacked for
the public to see? And I know this is an issue, David, we have talked about with some of the CABG work, for example.

But it is just something for us to think about because it does come up.

MS. JOHNSON: Okay, so -- oh.

Yes?

DR. WRIGHT: Just another contextual question. Is there any connection between our discussion and thinking about composite measures in terms of where the data comes from, i.e., electronic measures, meaningful use, that whole connection?

DR. BURSTIN: Yes, so we work really closely with the Office of National Coordinator. I am actually on Quality Measures Workgroup as well. So this has been an issue that has come up. I have not seen any composites come forward as part of that. I think it has been pretty hard to get the basic measures put forward. And I think it will be interesting to see how that comes
forward. I have not seen any work yet that moves towards composites.

But one of the things that keeps coming up, which is interesting, is some of you may have seen that as our criteria have gotten tougher on evidence and testing, in particular, a lot of the measures that were endorsed in the last five years or so are actually not making it through maintenance.

And one of the concerns oftentimes is it is a process measure. It is far too distal from the outcome measure. On its own, it doesn't work. But there has been, I think, some interest in saying can you potentially move towards an all-or-none approach of saying if these are all the right process steps that should happen, can those move to be compositied into something that becomes more of an all-or-none? I don't think we have a sense of how that is going to play out in an electronic environment yet.

DR. ZASLAVSKY: Why would it be an
all-or-none? Why would the composite have to be all-or-none for something like that, as opposed to any of the other methods?

DR. BURSTIN: You are going to get into this in a big way. For some of these examples, some of the individual process measures are at such high levels of performance, that otherwise there is no discrimination has been one of the concerns. And it may not be that that is the right approach. Maybe it is just time to just look at the outcome and skip the process measures completely but that has been one of the concerns.

MS. JOHNSON: Okay, so back to our issues. And I think we have articulated most of the things that are going to come up on these slides but one is distinguishing between instrument-level composites versus performance measure composites. So we have already talked about that in terms of the CAHPS measures. The measures that have come
through have been inconsistent in terms of implementation of guidance and forms, both again on staff side and committees that have to evaluate these measures. And of course developers who submit measures.

           All-or-none measures just don't seem to fit the additional analyses that we have indicated is necessary for composite measures.

           Sometimes developers either don't identify their measures in composites. Sometimes they do and they are really not. Sometimes they just don't want to use our composite form. So in that case they wouldn't be answering the questions that we ask in terms of evaluating composites.

           And also the thing that has been problematic at least internally is our composite form has only recently been implemented for online submission. So back in the 2008-9 project we came up with the criteria and then various submission forms
have been created but it has only been
recently that has been kind of put out so we
can grab data and store it electronically.

    So that is part of the reason and
I can't tell you specifically how many have
gone through maintenance. I just don't know
without having to go back to our paper
records. So a little embarrassing there.

    We also have had difficulty
applying the requirement and all of these kind
of merge together some of these issues. The
requirement that individual component
performance measures be NQF-endorsed or meet
all criteria. So what does that really mean
and how we apply that to all-or-none measures
is one of the questions that we have.

    Part of our guidance right now for
composite measures is that it is pretty easy
in a way if the components are NQF-endorsed
but I will read you from the previous guidance
document. "A component measure might not be
important enough in its own right as an
individual measure but it could be determined to be an important component of a composite."

So what does that really mean?

Some folks have interpreted that as not needing to meet our importance criteria, which includes impact evidence and performance gap. So that is our language right now but it is a little unclear what that means, unclear for all of us all the way around.

DR. DE LONG: I have trouble envisioning an actual example.

DR. BURSTIN: Go to the next slide.

DR. DE LONG: Oh!

DR. BURSTIN: So did I, so I added the next slide. So you are in good company.

MS. JOHNSON: Yes, so Helen will be helping us out on examples here.

And then finally on this slide, evaluation of the components themselves are challenging. And you know, you would think that would almost be an easy part of it,
easier than the scoring but sometimes the components are not endorsed as stand-alone measures. Sometimes they are competing with other endorsed measures and sometimes they are not harmonized to other endorsed stand-alone measures.

So with that, let's go to some examples and let Helen --

DR. BURSTIN: Sure. So I just pulled up three examples that I thought might express some of the issues we have had with components. And these are not necessarily a systematic review but the ones that really jump to mind for me, at least, in terms of the ones where we have had issues. So the first one is a measure that was put forward and endorsed by CMS, which was a 30-day post-hospital discharge care transition composite. It was actually a measure of three components, the first was a previously endorsed 30-day readmission. This was done for each of the conditions of CHF, heart failure, are the two
we got and pneumonia. That was already
endorsed. They then included an emergency
department visit and an E&M visit, a follow-up
physician visit. They assigned scoring so
that it was minus four points for the
readmission, minus two points for the ED visit
and plus one for the E&M. And at the time
this measure came forward, that was a logical
sort of compilation of thinking through what
a transition composite might look like. There
was some concerns about the waiting which were
really done, expert panel seemed logical,
readmissions are worse than ED visits but
again.

But the biggest issue was the
concern that it wasn't clear that the ED visit
itself really did capture a lot of the
exclusion of understanding the severity of the
ED visit. Was it appropriate/inappropriate?

And then concerns raised about the
E&M visit component, which was that it
completely excluded, for example, home visits
by nurses. There is no E&M code. It may be perfectly appropriate transition follow-up care. At the end of the day, though, the thought was composite itself was a really rich conceptual concept and it was reasonable to go forward but we did not, the committee did not endorse the ED visit and E&M visit components and we indicated back to CMS those are important. We would love to see measure that actually appropriately capture follow-up visits and ED visits but these probably weren’t ready for prime time as a stand-alone. But we thought as part of the composite they made sense.

Now I will tell you when that measure got to the Measures Application Partnership, there was a great deal of concern about how could you possibly have a composite that has components within it that were not endorsed or you didn’t think were appropriate for endorsement. So to me that was one example of the kind of issues you have
encountered.

Did you have a question?

DR. DE LONG: I have a comment more than a question. I have heard a lot of buzz about the 30-day readmission measure and I don't know if it is appropriate to talk about that today. This is just an example but I have heard of vignettes were patients were sort of pushed off or potentially pushed off and not readmitted so that they wouldn't count against the 30-day readmission. So that is a potential downside of some of these things that we are discussing.

DR. BURSTIN: All are not specific to composites. And certainly Sherrie knows this well, she chaired the committee.

DR. KAPLAN: I'm getting a little bit lost in the purpose of kind of the exercise here. If for me, from a measurement science standpoint, if you have got a complex construct under the microscope, then you probably, it is a multi-dimensional complex
construct, that is when you get into the position of having to create composites of thing that you are measuring, like math is a complex construct. There is algebra. There is geometry, there is calculus, blah, blah, blah. And each one of those things needs to be represented correctly.

Whereas you are counting discrete events or episodes like you are trying to measure maternal mortality, you count mothers who died. You know, it is not as complex a measurement exercise. It still has all the measurement issues associated with it, precision and validity, but it is a different exercise than trying to measure complex constructs. So if you are trying to measure quality of care for the whole hospital, that is a very different purpose that you are trying to accomplish with the measurement task you have in front of you than if you are trying to count bodies, you are measuring mortality rates for hospitals.
So just to sort of -- each one of those things is a different category optimal vascular care is a complex construct by definition.

DR. BURSTIN: Again, the intent of this was to just really not to get into the details of the measures but just to show you some of the issues you have so when you get to the guided discussion with Liz and Patrick, some of this will make hopefully some context will be helpful here. So that is just the first example. Again, specifically brought up this example because it was people view it as a complex construct that made sense at some level but didn't necessarily think the individual measures rose to the level of being endorsable as stand-alone measures.

The second example here I included specifically because of harmonization issues. So this is the optimal vascular care measure that I mentioned that comes out of Minnesota Community Measurement, which includes these
four components of LDL, blood pressure control, tobacco-free status and daily aspirin use. And part of what the committee wound up doing was harmonizing it and actually they did harmonize to the blood pressure control level that we already have for the individual stand-alone measure. But for example, there is no measure of tobacco-free status. There is a measure of offering help with smoking cessation and counseling but tobacco-free status is not a measure. Daily aspirin use is a measure only in a claims-based measure we have got which is somewhat problematic because it is not often on the med list but hopefully we will be going forward.

And so at the end of the day, this was endorsed as the composite and yet submitted on a single form as an all-or-none composite. So the committee never actually had the chance to say should any of those individual components go forward but the thought was at least at the end of the day
please have it harmonized to the individual measures we have already got, which is what we attempted to do around LDL and blood pressure for example.

DR. CHASE: Just to make a comment there because it struck me as you went through that that one of the challenges of why this can occur, too, is when you put together a new composite sometimes the individual components have a different basis so when they were constructed. So the denominator might be slightly different or so forth. So sometimes I think that is what is driving -- what we may run into and we need to be able to recognize that when you actually implement these sometimes you need to construct them differently because you are trying to get -- your construct is different. It is trying to get at something else than the individual component was.

DR. BURSTIN: Right and that has been a challenge in terms of the denominator
for all-or-nones being different than the
denominator for the approaches that often are
taken with the composites of weighted
measures.

And the last measure I put up
there just as an example is the patient safety
for selected indicators measures that AHRQ had
put forward. And the reason I put it up there
again is there are several measures in here
that as part of the initial evaluation were
not endorsed and yet thought as a general
construct and this was done by the first
committee, that they were appropriate for an
overall sense of patient safety, even if they
didn't feel like they were necessarily
measures that would stand alone.

So for example there is a measure
there about selected infections due to medical
care and concerns that well we have already
got measures from the CDC's National
Healthcare Safety Network, NHSN around CLABSI
and others and do we want claims-based
measures that would compete and not
necessarily agree with what people consider
the gold standard?

So again, just give this to you
more as a sense of these are the kinds of
issues we have encountered. So as we go
through the more formal discussion with the
chairs, I think you will have a sense of at
least what some of these terms means. Because
it is a confusing space for us.

DR. ROMANO: I think part of the
issue there, too, is with the reliability of
the individual components. So one of the main
reasons for constructing a composite is to
extract information from multiple measure that
may be relevant to a quality-related concept.
And to that end, some of those components may
not be able to stand on their own merits in
terms of having sufficient reliability for
public reporting at the provider level.

And given that NQF has kind of
raised the bar there to say that all the
measures have to be sufficiently reliable for recording at the provider level, that is the main reason to create a composite is to enhance that reliability.

DR. BURSTIN: And that is something I think we would love to have more discussion about is understanding that leap of saying it is not okay in an individual measure. When you put them together it increases reliability. It something I think we will need to spend some more time on.

Did you have a comment on that Dave?

DR. SHAHIAN: So could one theoretically then have an NQF-endorsed composite, none of whose components would pass muster individually as an NQF-endorsed metric and is that what we want?

DR. BURSTIN: We haven't had any. It is a little hard to imagine. But again, I think that the one place it could be is potentially as an all-or-none, where there may
be that the elements as constructed might be
slightly different. They may be the same
concepts but again because of the denominator
issues they may look somewhat different.
I mean we didn't, for example,
endorse the individual measures under optimal
vascular care, they were not submitted as
such. It was an all-or-none. And yet it was
interesting because when it actually came up
as part of the, I believe it was the ACO
payment. They really loved this measure and
I believe chose it but wanted to use the
individual components. And we are like,
actually we have never look at the individual
components of this measure. It was submitted
as an all-or-none.
So these are tricky issues for us.
I don't know the answer to that, David.

DR. SHAHIAN: I mean I think we
all know that we are dealing with a
proliferation of measures. And I am a little
concerned about the concept well for the
composite we need to have a slightly different
measure of the LDL or a slightly different
smoking measure than we do for this endorsed
NQF measure.

Then we have got people trying to
deal with 15 different smoking measures. That
is just the wrong direction to be going, I
think.

DR. ROMANO: I think we will come
back to that question a little bit more in the
discussion.

DR. BURSTIN: Yes, definitely.

MS. JOHNSON: We've already talked
somewhat about the evidence for each
component. Again, they may not meet our
updated guidance because it has gotten a
little bit more stringent. But even for
component measures on our current composite
measures where they were previously endorsed,
it depends on when that endorsement happens.
So as the individual measures come back up
from maintenance review our evidence guidance
again has become more stringent and those may lose endorsement.

Another problem that we have noticed as we have gone through is just asking about the purpose and quality construct that is part of our questions that we asked about composite measures. Those aren't always adequately explained or maybe even simply explained. It is sometimes hard to understand just what the quality construct is for some of these measures.

And then in terms of the measure specifications, they are often insufficient. Sometimes because they are just incomplete, things just weren't answered.

Often though, things are answered but it is hard to understand either what they did or maybe why they did it and then difficult to evaluate the analysis. I'm not sure we can do a lot about the last part because these are just complex analysis and not everybody has the statistical knowledge to
be able to understand every little thing but these are some of the issues that we have run into.

DR. DE LONG: I was just going to ask if this is an issue for us. I mean the incomplete and difficult to understand should, I would think, be turned right back. Now, the analyses could be a different issue but it seems that we wouldn't entertain something we couldn't read.

DR. BURSTIN: I think the issue is yes, we do try to work with measure developers, try to get information as complete as possible. I think sometimes the problem is it is actually hard to understand what we mean. So part of what our current efforts have been around our new work we have been doing around process improvement around our business development process and just have been piloted a two-stage process has actually been coming up with guidance that says what does good look like. What does this actually
mean? And I think part of what is also really
not clear for what we have asked people to
submit around composites is what is required.
So that half the time people may submit
something they think is complete but through
our lens it may not be. Or it comes to the
committee and they are like well that is
wholly inadequate. Well, it kind of meets
what it says on paper. So I think again,
being able to clearly say to developers this
is what is required, this is what good looks
like, I think will help us all a lot.

So we need your help in making
sure we are really asking for just what is
needed and nothing more because again, a good
number of you have submitted to NQF, it is not
an easy process. There is a lot of
information and so we don't want to ask people
to submit a lot of information committees
won't ultimately use. So you want to make
sure you are really honing in on what they
need to submit and the right way to get it
evaluated.

DR. ZASLAVSKY: Can you give a sense of how much of the work you are doing with these measures, typically these composites is really going through the components and how much of it is looking at the way it is composited? This seems to be just emerging as an issue that the evaluation of the components is becoming a major problem in itself and then thinking about whether the way you evaluate them is different.

I could see how, this may be a separate point, that it is a big difference between saying this is a component that wouldn't stand by itself because of a variance and lack of reliability and saying it is a component that wouldn't stand by itself because of bias because it is measuring the wrong thing or it would unfair to some institutions or something like that.

And I am not sure how much of the work is going to which side of these two
problems.

DR. BURSTIN: I think it really depends on the measure. It is very measure-dependent. I think there is ones when it has been much more about the construct and there are times when it is more about the components. And I think our goal is to figure out what that right balance is.

You know, for example, we just evaluated a very extensive perinatal maternal and child outcomes composite really nicely done. Some of the individual components were very exciting. It wasn't risk-adjusted. And at the end of the day 90 percent of the discussion was about the construct of risk adjustment for an outcome measure like that. And actually not as much about the individual components that everybody agreed were important but not risk-adjusted.

So it really is very, you have seen one composite at NQF and you have seen one, I think is truly where we are.
So the more we can try to make sense of that, the better.

DR. BIRKMEYER: I think with regard to that last bullet, I think there is a couple of issues. And I think the easier one to deal with is just the problem of applications that just aren't clear or are incomplete. I think the tougher challenge and one that value judgments will have to be made are with regards to the more complex statistical or econometric-based of composite measures that may be as clear as they can possibly be but couldn't possibly be evaluated or couldn't be replicated by a large majority of hospitals or users.

And that is where I think, that and composite measures are just fundamentally different from the usual business of NQF in evaluating individual indicators.

MS. JOHNSON: Okay, and just a little bit more information because we thought you might want to know. We have had six
composites that are not currently endorsed and
I just wanted to walk you through some of the
reasons.

One is lack of variability and
overall high performance. This came through
the central line bundle composite. So that
one, basically performance rate was very high
at 95 percent. So there was little
opportunity for improvement and that is why
that one went down.

There was lack of evidence
supporting components on an all-or-none
measure. That one, an example of that with
the ventilator bundle measure. And what
happened there, it was actually withdrawn by
the developer and because of lack of strong
evidence to support the measure focus, the
current national effort to define ventilator
complications and also I think the developer
may not have intended the measure to be used
for public reporting. So again, lots of
reasons the developer decided to withdraw that
one.

Patrick could probably talk to why
one of the AHRQ composite measures was
withdrawn. I actually couldn't find that
information but again, that one was withdrawn.

DR. ROMANO: We actually don't
think it was withdrawn. So that is a separate
conversation.

MS. JOHNSON: Oh. So we will get
together offline and try to figure out what
happened with that. Maybe I just have the
wrong one.

The component performance measures
were not endorsed and did not meet criteria.
Helen already gave you that example. It was
the perinatal adverse outcome index. And
again the problem with that one is the outcome
measures, the components were not risk-
adjusted. So they did not meet our criteria
around scientific acceptability.

Composite measures included some
performance measures that lost endorsement or
missing data had a substantial impact. That happened with the composite measure of hospital quality for AMI. That one was a new measure that we evaluated in 2010 but it had a smoking measure that was no longer endorsed by NQF and apparently there was a lot of missing data that was handled by imputation using national means. And that was just not felt to be good enough to pass that measure. And then finally some component measures that were more representative of quality of care were not included in the composite measure. That was another CMS measure of hospital quality for heart failure and that measure specifically didn't have components that dealt with beta blocker use, better discharge measures in cardiac rehab. So the committee evaluating that measure felt that all of the right components just weren't in the measure. So that is why that one went down.

So back to Helen's point, every
measure is different. Every composite measure came in differently and possibly went down for many different reasons.

MS. PAGET: I think you have already answered my question but we then make the assumption there was no overlap. So each non-endorsed measure is not endorsed for a different reason. We are not seeing patterns, for example.

DR. BURSTIN: I don't think we are seeing patterns other than I think some of the issues we have already brought up. But I think also the patterns we are seeing is this issue of the components within a composite. So people may sometimes like the construct but not the components. Or people sometimes like the components but the contract is not risk-adjusted. So I think it is kind of all over the map but interesting this last one as well, you know, we are currently looking at a colonoscopy composite that is being reassessed by our GI committee, for example. You know,
one of the concerns was those are some of
those indicators as part of the composite were
really important and useful. But at the end
of the day, the GI folks in particular thought
the best way to really look at the quality of
a colonoscopy is your adenoma detection rate
and all the rest of it was kind of on the path
towards getting at what is really important
and that wasn't part of the measure.

So often times I think we are also
hearing this issue of is it really capturing
truly the quality construct you care about
that the composite is allegedly trying to

DR. DE LONG: Can I ask? On that
last point, does the composite have to be
comprehensive and where, for example the heart
failure one, apparently the components are all
individually endorsed and are used as stand-
alone measures. Is that correct?

DR. BURSTIN: Yes, although many
of those measures were topped out, really,
really high levels of performance. A couple of them so topped out or in one case the smoking measure had been reviewed by NQF and we didn't really believe it was a valid indicator of smoking cessation in hospitals. It had become sort of a complete checkbox measure. So it was not endorsed. Most of the others were topped out. And so again this issue of what do you do with a composite that essentially gives you information that is not a whole lot different than the individual measures.

DR. DE LONG: Okay, my question was a little different.

DR. BURSTIN: Okay.

DR. DE LONG: In terms of, for example, beta blockers for heart failure, it was not included in the composite but presumably it is a measure that is being used on its own already endorsed. Is that not the case?

DR. CHASE: Well I would have an
opinion on that, which is I don't think a criteria should be the reviewers could come up with other things that could be included in the composite because that may not be the purpose of how the measure was constructed. And our example in diabetes care, for example, there are some things about treatment that aren't in there because that measure was constructed around a risk reduction. So as you know in diabetes care, the other things like the foot exam that might be really important for care but it wasn't what the purpose of the measure was. So I think it would be dangerous to sort of have committees saying we would like this to be more inclusive when that may not have been the original intent. And while I hope we get to this as we talk later I saw on the construct of this, composite measure, people who are presenting those should be able to articulate what is the purpose of the composite. Why did you
composite things?

And then the test should be is it doing what you intended it to do. Which may not always be everything about a particular case. We weren't trying to construct this to try to cover everything about a condition.

DR. DE LONG: And I still want to go back to for example Dave said that there have been a proliferation of measures. And I don't think every measure that is being used is a composite.

DR. BURSTIN: No.

DR. DE LONG: There are single measures.

DR. BURSTIN: Right.

DR. DE LONG: And if they are being used on their own, for example beta blockers after a heart attack, you wouldn't necessarily want to double count them as including them also in a composite, necessarily.

So leaving out some measures that
are already endorsed on their own and felt to
be sufficiently important, it seems to me we
have to decide. Do they go in a composite or
do they stand on their own and not have a lot
of overlap there?

DR. BRATZLER: Patrick?

DR. ROMANO: Go ahead, Dale.

DR. BRATZLER: Yes, thanks. The
other things that I think has to come into the
discussion, we will probably talk about it
some later, is the issue of unintended
consequences of the composite.

So Helen gave the example of
composites that were built with largely topped
out measures. But particularly if you have a
composite where you have a number of
relatively high rates of performance, you have
other components of the composite that have
lower rates of performance the consequence of
the way to improve your performance rate is to
focus on those aspects of the composite that
have low rates of performance or the biggest
denominators, which may or may not be the most important components of care that might result in better patient outcomes.

So there are a lot of unintended consequences that can come up when you group measures together.

DR. ROMANO: Well I am just mindful of the time and I don't know -- we do have some questions that we want to discuss specifically. But if there are additional questions related to Karen's presentation or comments on Karen and Helen's presentation, we should get those on the table now.

If there are more general questions related to our discussion of conceptual framework and so forth, maybe we could hold those for a minute or two.

DR. KAPLAN: This is just sort of wrapping around the final points that were being made, which is again, what are we measuring? What are you measuring? If you are measuring discrete events like beta
blockers following something like that, you
are measuring a discrete event. And the
measure is good for that purpose and only that
purpose.

If you are trying to use that
measure like two trains left Chicago traveling
at a distance of and you are trying to use
that to measure something else, that is when
you get into trouble and these complex
constructs have to be defined.

If we don't define what are you
trying to measure here and you are using that
to measure all of the quality of diabetes care
and there are things like foot exams that are
relative to that, that is the construct you
have got under the microscope. So specifying
that has to be really critical when you are
talking about these composites.

DR. BURSTIN: I think part of the
challenges we have seen is sometimes people
going so focused on the components that they
sometimes get lost and don't get to the
detailed, this is the complex construct. Not all the time but sometimes it has been a struggle to say, you know we will ask for quality construct and what they will do is a recitation of the components. Like no, no, no, what is the construct?

DR. ZASLAVSKY: I have a question. This is sort of more about how NQF sees its role.

Suppose that you have three different groups come in either at the same time or at different times and they want to develop a composite or get approved a composite for a certain construct and they are even using these same components and one of them says we are going to equally weigh them. Another one says we are going to use factor weights. Another one says we have some kind of criterion-based regression and we have done it on some dataset that we have and here is another set of weights. Is it your role as NQF to pick one of these or to let a thousand
flowers bloom, or the first one past the gate
is it and the next one has to overthrow it?
How do you see your role in that kind of
situation?

DR. BURSTIN: It is a tough role
but it is one people increasingly look to NQF
to really try to sort out. Again, the
proliferation of measures David mentioned the
cacophony some might say, there are so many
measures at this point that are slightly
different, slightly competing with each other
that actually have done a lot of work,
particularly in the last yeartrying to figure
out exactly how we assess related and
competing measures which is what we call them,
when it is actually trying to come up with
some sets decision rules about when two
measure can move forward. For example, the
same measure harmonized different settings of
care or different data sets, potentially could
still move forward as long as they are
harmonized.
But when things are truly competing, we really do try to have as much as possible committees try to go through them and say which of them is actually best in class or superior. It is often very difficult to do. We don't have data runs for example to be able to go back and say head-to-head, if ideally you could have this measure and this measure of construction on the same dataset, it is often an impossibility.

But we do try to, as much as possible, pick what we think is the best in class. And if that is not possible, sometimes we will move both forward. I mean, one example is at the time of the diabetes project in our outcomes project about a year or so ago, we had the all-or-none composite on optimal diabetes care from Minnesota Community Measurement. We also had the weighted composite out of NCQA. At the time, the committee tried very hard and at the end of the day could not say one was necessarily
superior. For different end users, one might fit better than another did and they thought the constructs were different enough. And again, we worked with the two developers such that the components within them were harmonized. Meaning that for the individual clinician or somebody being measured, the blood pressure control was the same. The A1C is the same. The lipids are the same. So you have at least harmonized on science but we can't always harmonize on the approach.

So for now at least we allowed both of those to move forward, recognizing different end users might use them. And hopefully, as we gain experience, we will have a better sense of that.

Sometimes the committees are much more clear. I mean, the cardiovascular committee just came down and could not have been more clear that at this point in time in cardiovascular care an all-or-none was the appropriate approach because these were things
that should absolutely always be done to move
the needle.

But again, this is where I think
we still see variability of cross committees,
which is why we wanted to bring you guys
together to try to have some guidance we could
share with the committees that they would
always act more consistently.

DR. ROMANO: I think this is
something else that we will tee up for
discussion a little bit later but I know in
the case of the AHRQ composites, we actually
offered three different weighting schemes and
said that different users may apply different
weighting schemes, depending on their
particular decision-making context and we can
talk more about that a little bit later.

But it ended up that one of those
weighting schemes became the NQF-endorsed
weighting scheme. But the other schemes are
still out there on the table for other users.

So it is an interesting question.
MS. JOHNSON: Okay, so now I think we should just go ahead and hand it over to Patrick and Liz to do the guided panel discussion. And I'm not sure how you want to do that. We have a list of questions that we hope you will address and I will just let you decide if you want to go through them one at a time or how you want to do that, Patrick.

DR. ROMANO: Well, I think obviously we didn't get a chance -- I'm sorry about the train but I think maybe what we will do is to start by talking about the conceptual model. And I think it has already come through in the discussion so far that most of us feel strongly that there should be some conceptual framework for a composite that the developer should be able to articulate a conceptual framework.

But the question is is this, there is a model one, model two that is presented. There are a variety of terms. So let's start out by discussing that, if that is an
appropriate framework. And then I think we should discuss a little bit about maybe what is not on the table today, what we are not going to consider as composites for the sake of the discussion the rest of the day. And then we can move into some of the other questions. Does that seem fair?

MS. JOHNSON: Yes and I do want to just make sure that we understand that we are not saying that XYZ measure is not a composite but just that we might not need to apply extra criteria to evaluate some measures. And that is how we try to construct those tables. So as long as we are good on that.

And I think the other thing about those tables is it was just our effort to put something out there. We don't want to get bogged down in taxonomy and all that kind of stuff but hopefully that might be a guide to help us at least know what we are talking about when we are talking about things.

DR. ROMANO: Thank you and I
really want to compliment the staff work that was done in preparation for this meeting. It was really a very impressive compilation of resources and organization of the key questions. So it really gives us a great foundation for this discussion.

So in this memo, NQF staff have sort of put forward these two general models that are described in the field and it is interesting I have heard before the psychometric versus clinimetric distinction but then I read David Streiner's argument here that this is a distinction without a difference.

On the other hand, then when I read his second paper, it was like well, he is really talking about the same thing. He talks about it in terms of scales versus indexes. Other people use the term formative versus reflective. Other people talk about whether the indicators are causing the construct or the construct is causing the indicators.
But really all of these, it seems like, are different semantic ways of describing the same fundamental distinction, recognizing that the specific approaches that might be used for evaluation are on a continuum really. And so with one model you might emphasize certain approaches, with another model you might emphasize other approaches. But let's just put that on the table for discussion. Do people find this helpful or not? Where should we go?

DR. KAPLAN: When Alvin Feinstein first cooked up clinimetrics he used to accuse those of us who were trained in a different discipline as worshiping at the altar of psychometrics. And so I have learned over the years to call it measurement science and it gets you out of a lot of touch calls with your clinical colleagues and your psychologically -- measurement science is really at the base of everything and it doesn't matter what is under the microscope. All of the principles
apply.

DR. SHAHIAN: And just to add to support to what you just said, I don't know who put this particular dichotomy together in terms of conception models one and two, but I can tell you that having the STS CABG composite as an example for concept two, that is not the way we thought of it. And in fact we spent probably over a year evaluating this model from a psychometric from a traditional psychometric perspective.

So at least that is the way I went about it. And I think most of the people on the team, Liz might want to comment, but -- so I don't really think it is appropriate to list STS as an example of the clinimetric.

DR. DE LONG: I think we are getting into semantics. As a matter of fact, on the phone the other day we discussed whether it really matters how the metric actually came into being. What really matters is what it reflects and whether it is
reliable, whether it could be reproduced and it can be used to improve quality.

So that may even relieve you from your having to take a part all the methods and put them back together if the developer can actually produce something that leads people like John's company to be able to implement them and they work --

DR. BIRKMEYER: I didn't say that.

DR. DE LONG: -- then maybe you have got a good measure.

DR. BIRKMEYER: I have got very mixed feelings about whether the single most important thing that we need to start with is that there be a conceptual model in place.

I totally agree with Sherrie that the most important thing is to be clear what at the end of the day that you are trying to measure and the need to work backwards and see whether that goal is being met why whatever the measure is.

But I think, and I am reflecting
my bias as a simple country surgeon, that kind of at the end of the day, there is a lot of measures out there that really get you to the end result of what you are trying to achieve and you don't actually know exactly why that measure was the best one to get you there.

DR. ROMANO: Well just to pose a practical question. So I agree completely with sort of banning the terms clinimetric and psychometric and focusing on measurement science and recognizing that this distinction between model one and model two is really a spectrum, perhaps, and not two bins that you have to go into bin A or bin B.

But to give an example with David's measure, so you explicitly looked at internal consistency reliability. But someone else approaching this same question might say that they don't care about internal consistency reliability because they are looking at five different types of outcomes or complications, if you will, of cardiac surgery.
and it doesn't matter whether those hang together. Some clinicians may do better at preventing A. Others may do better at preventing B. But from the patient perspective, they are all bad things.

And so from the patient perspective, it makes sense to add them up in some way. So how would you respond to somebody else who said no, we are going to construct this measure without regard to internal consistency reliability because we are approaching it from a different perspective.

DR. SHAHIAN: I guess my personal bias is that I guess I am somewhat of a traditionalist. I would probably, I think tend to be more along Sherrie's way of thinking.

I think if we haphazardly combine things just for the sake of combining them and don't know whether item B adds value to item A as a stand-alone in evaluating the
underlying latent construct, or if we don't -- if they are totally redundant, these are important considerations. And I think frankly there is all too much of this sort of just haphazard pick a bunch of things, and put them together and call them a composite.

I would rather, personally, see a more traditional measurement science approach. That is my personal bias.

DR. ZASLAVSKY: I have to say that I have been through this argument over and over in development of the CAHPS surveys. And I can't help agree with what Patrick is saying about there being these different perspectives that do lead to different decisions, especially about, not necessarily so much about how you composite once you know what you are compositing, but about what you need to put into the composite in the pursuit of a high alpha leads you in the different direction than sometimes what you will get if you are trying to group things that you think
are similar more in the way that Patrick describes as not necessarily being empirically related but being conceptually related or to a similar criterion outcome variable.

So I would love to be able to not have this be an argument but I think that these do sometimes pull in different directions.

DR. KAPLAN: This is one of the times when I think Alan and I can actually reach common ground about perspective because I think getting the tyranny of psychometrics, if you will, has come out of real traditional latent construct. I have things I can measure about you that get to your IQ, et cetera, et cetera. And sometimes we shoot more in techniques for getting to reliability.

Things like physician-level reliability, well how do you tell whether using a set of things about diabetes care, for example, whether those measures are a reliable estimate of physician's performance. The sort
of traditional internal consistency of reliability may not be the right approach for that particular measurement task. And then you need to look at intraclass correlation or has the physician got a thumbprint. Is the physician behaving consistently across patients within their practice? So the technique and approach may be different for the different measurement task we have at hand.

And I would not like it or maybe it is -- you know, I hate to test reliability but maybe that is the right approach. There are all other kinds of techniques that you can use to get to capital our reliability and I don't want us to -- I wouldn't like to see us get bogged down in one "disciplinary perspective."

DR. SHAHIAN: I think each of those methods that you are talking about is a measurement science approach. I am not wedded to Cronbach's alpha but all the things you
mentioned have an empirical basis.

DR. KAPLAN: Absolutely.

DR. SHAHIAN: That is all I'm saying.

DR. DE LONG: And Alan I think was the first person to say the word outcomes. And it seems that what we really are driving at is something that improves outcomes and it doesn't matter so much how much it was internally consistent or whether it all reflects the same thing. I do agree, they can reflect different components that are not necessarily tied together.

MS. PAGET: Just a couple questions. I think I brought this up on our call but my question to the experts here in measurement, do we see this same kind of debate or tension exist in other industries, most specifically I am wondering about educational testing that uses a lot of composite measures. And then my second question I guess maybe is for Helen is how
important is it that this group and NQF fall
down somewhere on this issue? I mean is it a
deliverable of this group to really state
something about the conceptual model?

DR. BURSTIN: I'll answer the
second question. I don't know that we have to
say that. I actually think Patrick's comment
earlier I think it was Patrick that we
potentially will be needing a spectrum. I
just think we have to have some guidance for
the committees to say what level is
acceptable. If they are both acceptable and
they are just variants of the spectrum, all
based in measurement science, I am fine with
that.

And I think even as the team was
putting this together there was a sense that
how much gray is there between these first two
models? I just wanted to apologize for
mischaracterizing the CABG composite. I just
wanted to have something to put together to
put in front of you because we do get very
different approaches that people bring to us. If either is fine, okay, good. I don't think we need to pick one necessarily.

MS. PAGET: Yes, just wanted a particularly educational -- does this kind of question exist or are they -- before, after, in the middle?

DR. ROMANO: Well it is interesting that you raised that question because that question really was what inspired Jeff Geppert's paper, which I think was part of the packet which is currently under review, which is a belief that in some other fields, particularly in financial services, there is a very strong emphasis on what is the decision-making context for a composite. So when a composite like the Dow Jones Industrial Average, for example, or the consumer price index is created, there is a lot of thought that goes into how are people going to use this to inform their decision-making and let's construct the composite in a way that produces
the right signal that encourages the right allocation for effort, if you will, the right investment of resources across different sectors.

So I would like to put that out for discussion for a couple of minutes to see because to me that gets back to the concept of different weighting schemes and why AHRQ actually offered different weighting schemes.

Because if the approach is, for example, we have an indicator that is called patient -- it is a mortality across multiple procedures. And it composites mortality, risk-adjusted mortality, for different types of procedures. And it was not endorsed. The committee was concerned that it was too heterogeneous because it was bringing together different types of procedures done by different types of surgeons.

But there could be a counter argument made that in the right decision-making context, it would be important to
signal what the hospital's overall quality was for surgical procedures. And that it might be useful for the hospital to understand something about what is driving that overall performance.

So that is a case where perhaps what we should be thinking about more maybe, instead of focusing so much on how these building blocks are put together, maybe we should be focusing more on what is the decision-making context. How do we want people to use this and is the composite constructed in a way that is consistent with that use?

DR. KAPLAN: Again it gets back to what you are trying to measure because educational settings and educational circumstances have cognitive performance often -- usually as their base and that is a different enterprise, measuring cognitive performance from behavioral performance which is usually what you are trying to get at when
you are looking at many of the measures of quality. It is the performance of something. And in case of outcomes, it often has somebody else's performance at its base. So behavioral and cognitive are two different exercises by far because those -- and now teachers' pay-for-performance by the way is trying to use students' performance on standardized testing, so I can compare the same test across students to reinforce teacher performance, as an estimate of teacher performance. But it has the advantage of being standardized because it is cognitively based and it has the advantage of being widely tested on a lot, a lot of folks so we know a lot about that measure at the student level.

But now we are looking at patient-level variables and we are trying to wind them up to estimate physician level performance and then we are trying to wind that up to measure clinic or institution-level performance. So we have got a different -- the behavioral
stuff is a very, very different measurement exercise. It still needs to respond to what are you trying to measure and all the things about that and are you able to do it consistently and are you able to do it with some level of accuracy?

DR. BIRKMEYER: So just to follow up with that, I totally agree with the point that Patrick just made that while being clear about what you are explicitly trying to measure, you can sometimes only answer that question if you consider it in the decision-making context and what types of judgments are people trying to make at varying levels of altitude. The challenge, of course, though, is that runs counter to the NQF mantra of we consider measures sort of agnostically with regards to their application and maybe like that is why composites may be different.

DR. DE LONG: I think there is another issue that was in one of the references you sent and I can't remember which
one. But there was an experiment where they took two different datasets and tried out five different methods. And there was almost no consistency in the results. The top performers for one dataset were by one method top and by another method bottom. I think when we look at why we are developing these measures we also need to be thinking about whether they hold up. Whether, for example when the developer puts them out there, they maybe had a big enough sample to split it and see if that measure actually performs the same way in the other half of the data. And I'm wondering if something like you are suggesting the reason this is connected to your comments is that maybe when you get something so broad as surgical mortality for a whole bunch of different surgeries, that really isn't -- once you combine it, you are not reflecting something that is stable. I don't know.

DR. CHASE: So kind of back to your question I think about do we -- is it
worth spending more time about these two
different models? I would be more on the side
of no. I am reflecting into what our purpose
here -- it would be a really interesting
discussion for us here but when sort of this
gets applied at NQF, I think we are trying to
get to advice for reviewers about what is
different about composite measures.

I don't think it is going to be as
practical when you get down to so what are --
unless and I couldn't find this about what I
would say. Oh, if you fall into this bin, you
have to do these things. And if you fall into
this bin, you have to do these. I think in a
couple of the other models we do, but the
distinction in the first two I was getting a
little bit lost. And maybe that is the other
ing how I would say about this is is
probably a lot of your measurement developers
and the reviewers might get lost by this
terminology, too. And then it is not going to
be helpful.
So I would encourage us maybe to move down the list and deal with what are some of the criteria around the other ones. But that would just be my thought here.

DR. DUNTON: If we wanted to divide up the world, --

DR. BRATZLER: This is Dale. I think I also agree with that.

DR. DUNTON: I think that rather than think about the model, it might be useful to think about a composite of process measures, versus a composite of outcomes. Because the question for the process measures is was optimal care provided. It could be all or none, a percent of the time, or something. But if you are looking at safety in terms of outcomes, it is probably less likely that you are going to get really reliable measures than you would if you are looking at was care provided. And so the measurement exercise may be different. The committees reviewing them would be maybe have different standards for
the measurement level, the scientific
acceptability criterion and I think that is
ingoing all I should say because I am getting into
territory where I --

MS. PAGET: So I guess I agree
with what I am hearing completely that not to
get hung up and I certainly don't have the
knowledge in which to kind of form an opinion.
But I do have to say that there are some
things about the Geppert paper that are
appealing to me. And one is the role in the
weighting of process versus outcome and the
whole concept of signaling that effort that
could be much more in tune to the necessary
versus unnecessary care and treatment.

And maybe I am reading too much
into it but I guess my question about that
conceptually is does that open up -- does that
kind of thinking open up an opportunity for
NQF to actually be endorsing more measures
that are outcome-driven because it has a means
by which you can balance these two process and
outcome? I know that I hear repeatedly the
movement toward more outcome-based measures
and I just, when I read this conceptually it
feels to me as though there are some themes in
here that might be important for us to
reiterate in a guidance or whatever product
that comes out of here. And that is one that
to me feels like it could be conducive to
where NQF is hoping to go.

DR. ROMANO: Yes, so I think what
we are hearing -- if I am not summarizing
correctly please stop me. But I think what I
am hearing is that we don't find this two-part
conceptual model terribly useful. I mean
fundamentally at the end of the day, the
purpose of this exercise, the reason we are
here is to provide better guidance to measure
developers and to steering committees to help
them submit composites and to help them
evaluate composites.

And so from that perspective, we
don't -- we want people to use measurement
science. We want people to use the appropriate tools from the armamentarium of measurement science but we don't know necessarily want people to -- we don't want to force people into a particular bin here based on this conceptual model because this may be an over simplification. This may be to some extent a false dichotomy. Am I correctly capturing what people are thinking?

DR. ZASLAVSKY: I agree with that as far as it goes but I think we shouldn't underestimate the importance of having a conceptual orientation in developing something. And the concepts that underlie that dichotomy are useful concepts and could make a real difference.

You might end up with several composites where there is a really clear conceptual argument for using one of them for public reporting and for using another of them for pay-for-performance and another one for internal process improvement.
And so asking people to make that part of their submission, you know to be very practical about it, not with a view just to putting things in a bin but to giving the reasons why, the way this was constructed makes it particularly good for particular purposes I think really should be part of the exercise of evaluating the composite.

DR. ROMANO: Thank you, Alan, that is perfect.

DR. KAPLAN: Well I agree because I think that if -- but maybe what is needed is give people some examples and some guidance. For example, if you are trying to estimate patients' experience with the doctor's communication, here is an approach that includes internal consistency reliability because I have a set of things that I am trying to measure, all of which I think measures a patient's experience of a doctor's ability to communicate with them.

If I am trying to measure diabetes
outcomes and I am trying to estimate physician's performance with those measures, here is what I need to look at. I need to make sure that if I am using it for physician performance, there is a physician thumbprint that I can show you that there is some -- doctors behave consistently and they differ from each other. So that is an example of that kind of evaluation.

If on the other hand I am trying to get risk adjuster for the total illness burden index, then is a patient with gastroenterologic problems likely to have cardiovascular problems, likely to have difficulties with joint disease. Probably not. So internal consistency reliability in that case wouldn't make any sense. So how am I going to tell if there is consistency across the things I am measuring in sort of a review of systems perspective, so I can get a composite measure that makes sense to me that estimates a patient's complex -- the totality
of a patient's complex comorbidities?

So you might want to cluster these things or American Board of Internal Medicine has no trouble with cognitive performance of physicians. They use it to accredit physicians. So if you have got a different kind of performance measure under the microscope, that has a different set of more like a cognitive performance psychometric approach. Maybe some examples by category of things, whether it is -- and those kinds of things might help people who are submitting measures look at what you are asking them to do.

DR. ROMANO: What makes it easiest I think for NQF staff and steering committees is if it is a menu. So you choose. You have composite type A. Then you submit A1, A2, A3, and that leads to this decision. If you choose B, then you submit B1, B2, and B3. But I think what we are saying is it is not that simple.
DR. DE LONG: I was just going to ask what are A and B.

DR. ROMANO: Well I mean A and B would be hypothetically Model 1 and Model 2 here. But I think what we are saying collectively is that it is not that simple. That it does need to be sort of written out. What is the concept that we are trying to measure? And what is the aim of that measurement? And then what are the appropriate tools for evaluating whether the measure is accomplishing those aims. Is that --

DR. KAPLAN: Yes, from the measure developers' approach.

DR. DE LONG: But I understood you to have different concepts, rather than conceptual model one and conceptual model two, that you listed three I think. And I thought we were going down the road of maybe there are different buckets that we could elucidate and then start with A1, A2, and A3 in terms of
what the requirements are. Is that not -- did
I misunderstand you?

DR. KAPLAN: Well I think what I
was trying to say is there are different
approaches you would use to reliability for
one purpose or for one construct maybe and
there are different approaches you would use
for another. So if you gave examples, so if
you have what is the measurement today -- what
are you trying to measure? How do you know if
you are doing that consistently across a
composite? What evidentiary approach are you
going to use? And then for the validity side,
are you going to be accurate? But it is going
to vary by what are you trying to measure?

DR. ROMANO: Can we predefine all
the relevant buckets or is that a task that
fundamentally has to be left to future
discussion in future steering committees?

DR. KAPLAN: Well I don't want to
dominate this conversation but I think that
you could probably give examples that would
elucidate that. For most people who are going
to come in with measures, you are not going to
be people who aren't at least some kind of
cogency with respect to measurement aren't
going to come in with measures to begin with.
So if you give them examples, here is the
kinds of approaches one could use in
estimating physician performance, in
estimating hospital performance, in estimating
this class of variables like if I am going to
add up all more mortality for the hospital,
what evidence is there that that is a measure
of the hospital's performance and is it
consistent across all subcategories of related
mortalities that you are trying to evaluate?
So I don't think you can do it in
a here is the Betty Crocker formula and you
are going to come out with a cake. You could
come out with a can of dog food.
And you don't want to end up
giving people -- being so rigid that you
really stifle because I think Alan is right.
In this case we are at an interesting sort of stage of the development in the clinical arena of creating these composites and everybody is all edged up about it. So I think if you tried to be rigid about it at this juncture, you would probably even stifle creativity. I don't think it is a good idea yet to kind of really lock people into one approach. If you don't shove a chrome box alpha in there, everybody's head is going to explode. I don't think that is going to -- that is even rational.

DR. SHAHIAN: No. There is a very wide spectrum, however. I mean they are trying to combine all risk-adjusted mortality rates or determining physician reliability and a thumbprint of a physician or a physician group is on one end of the spectrum.

Getting the X elements of a central line bundle or a ventilator bundle, which makes no pretense whatsoever of having any empirical basis whatsoever is at the other
far end and we are trying to encompass all
those within something we call a composite
framework. And I am wondering if the latter
really even belongs in this evaluation
framework.

DR. ROMANO: Well that is our next
topic for discussion.

DR. GOLDSTEIN: Yes, I guess
thinking about this, and I keep thinking about
the composites we use in my division, so we
have the CAHPS ones which are pretty straight-
forward. It is doctor communication or access
or things like that. But we also use
composites and we have never submitted this to
NQF so I'm trying to think how it fits in
there.

For example, for Medicare
Advantage we have an overall rating that
combines 50 different measures, some
individual ones, maybe NQF-endorsed, some not
and there are lots of judgments we have made
along the way with weighting of measures and
how we do the calculations and all of that. So I was kind of struggling listening to this how compositing differs a lot for what we do on our surveys is very different when we are coming up to evaluate a provider and coming up with this overall rating to say this is a high quality provider.

I'm trying to figure out how that all fits together. And when you get submissions, you are going to get at that wide spectrum. And what you may do for a survey is going to differ from what you may do for a clinical measure.

I remember early on for the composite forms for NQF, and I can't remember if it was a home health survey or one of them, we just struggled filling it out because it was really made for a clinical measure and not a survey measure. And I think it changed over time. But you may need to think of different things for different types of measures, whether it is survey, whether it is process-
type measures, or whether you are measuring kind of at that broad level of the quality and performance.

DR. CHASE: I like this discussion coming in here again because I was trying to get to I think when you put together a composite again you should have to say what was your thinking about how this gets to be used.

And so I take our example with diabetes. It was an interesting one because you have to put that in the context of the reason why we did that composite originally was because when you looked in the gaps in the community, it was completeness of care. You could see there were these five things that were in the guideline and over here they did these three well and over here they did these three well. And there wasn't any rhyme or reason of how that happened, other than just sort of practice how it had rolled forward.

So that became the reason why an
all-or-nothing made sense to bring those
together. It wasn't about now people argue
against it around well it doesn't really
reflect reducing risk in patients. You know
when you do all-or-nothing, it doesn't help
you when a patient is at 7.1 for an A1c
compared to somebody at 9.0. You get the same
credit to bring both of them down to 7.0, so
it is no good. And they say well that is a
different measure. You can do a different
composite that would be about risk reduction
and then it would need a different kind of
testing because then how you construct a risk
reduction measure is going to have different
reliability than an all-or-nothing.

So I wanted to tell that one
because once I go through and say that is a
really important thing, then I think there is
another kind of measure that people are going
to bring to NQF, which was to the earlier
point about people are feeling like there is
way too many measures out there and it is too
hard for consumers to understand. So I am trying to get back to your point which is people are going to come and say well we want to do a composite prevention measure because it is silly to give people 12 different individual components and all they are trying to do is make it simpler for people to see how much prevention there is.

So is that good enough? Do we get to where we say we need a construct but at the end of the day one construct may just be because it makes it easier to use and are we okay with that or is that -- and maybe there is some science around what is acceptable in making it easier to use by just glomming some things together.

So I am just curious what people think about. If we are going to set this up, are we really giving any differentiation criteria for a review panel?

DR. DE LONG: I wasn't going to posit an answer to that but I am curious about
your distinction between quality of care and risk reduction. Maybe a naive question, I'm not sure I understand. I would think the purpose would be risk reduction. And your quality of care should be totally in sync with risk reduction.

DR. CHASE: Yes, they both are dealing with risk reduction but I am saying you could construct them in two different ways for a different purpose. If you are trying to show patients this is what you need to completeness of care, you don't really care that the doctor is good at it with 90 percent of the patients. You want to know about it for yourself and again, if that was it, as opposed to this issue around now maybe we have a different goal in the community around -- we have pretty much got people accepting here is the guideline and trying to implement it consistently across the population. Now the effort is let's be as efficient with resources as possible. We are trying to get to what for
the given effort how much risk are we reducing in a population?

So that is why I thought it was really important to when you are bringing a composite measure forward you should be talking about -- you should have to articulate why the measure is being -- why it is a composite as opposed to the individual components.

DR. BURSTIN: Just to follow up on Jim's point and I think also to touch on what Liz was saying, as well, we have heard interest, for example, from CMS of saying can you take a whole bucket of all these measures that live on Hospital Compare that are cardiac or a whole bucket of these measures that live on Hospital Compare about something else and just create composites? Would that be easier?

And so at the end of the day, even if we get away from these conceptual models, I still think it would be helpful, I think to Jim's point of at least seeing if measures
came forward that took all 50 and said we have
taken all 50 because it is what we have got an
we think it is a pretty comprehensive view of
cardiovascular care and hospitals. What would
be required, other than saying we took the 50
we had in hand and here it is?

And so again, I am still struck by
Patrick's earlier point about this being a
spectrum. And I almost wonder for the
afternoon if it might be, if it doesn't take
us too far off base almost useful to kind of
almost create the anchors on the sides of the
spectrum and see if there might in fact be
different kinds of questions to at least give
a sense to developers of what would be
required to put forward. If you are bringing
a measure forward that truly is the 50
measures I have already got in hand in a given
topic area, what kind of additional analysis
is required for the construct to say it is an
acceptable composite versus the detailed
analytics David did to create the CABG
composite, which I think is probably the other anchor or the CAHPS composite. It might just be something to think about.

DR. KAPLAN: The only problem with that Helen is you end up sometimes adding up apples and airplanes and you can't. It has to be driven at the base by some construct that is clinically and from the -- it is not just, those of us in the measurement science arena can do a lot with the empirical stuff. But at base it has to be driven by the people who are -- what is it you are trying to measure? And that is always definitional. That comes from in this case it is probably the clinical and the health services community. What are you trying to measure? And then are these a good reflection of it? Because you can add up anything. But then the question is does that make any sense to anybody or are you adding up thing that are very, very different and very, very divergent and although you can certainly add them up and create an index. It makes
absolutely no sense to anybody.

DR. GOLDSTEIN: I guess I was just
going to add to, and I don't know how this
fits in the NQF process, NQF process tends to
be kind of a long process. So if I think of
our like our health plan program, we
reevaluate every single year the measures
included in our roll-up. So if we have
measures that are topped out, they go off,
quickly off. And plans don't like that
because they say oh, we improved in that
measure. That is a measure we are doing well
on and you see a mess and took it off. But we
are reevaluating and we keep adding new
measures to it in areas where we think we are
missing measures. So it is a really a dynamic
process and when the data goes live in October
each year, as late as August we may make a
call. This measure, one of our 50-something
measures there is an issue with it this year
and that will come off of it.

So it is a very, very, for that at
least, a very dynamic process. And when you think of the NQF process, it works very well for like the CAHPS measures or measures like that that don't change basically year to year, although there are differences and reliability and things year to year that we look at for the CAHPS measures and we make a decision to take. We have, I guess one good example from our CAHPS prescription drug plan survey that we are seeing now. You know, very little differentiation across plans. So we are going to actually remove it from our rating system.

But how in this process are things reevaluated really quickly and have more dynamic process with NQF?

DR. BURSTIN: I don't want to get us off track but we do have a process for annual updates, as well as ad hoc reviews anytime a measure changes. We are doing one on Monday actually. So that is part of our process. We have tried to be more nimble.

To me what it really speaks to is,
I think, this issue of we have lots of discussions with CMS over the years is this idea of are composites more conceptual frames of which you pop things in and out or are they actually grounded measures? And I think we have had this debate at times saying it is very hard for those being measured to say it is a conceptual thing, we will pop things in and out. And yet I understand that the reality of administering a program is things do sometimes change. It is an interesting issue. I'm not sure it is specific to this committee but it is one we have heard a lot about. Well can't I just say I am using endorsed measures and I am creating a composite really essentially what the STARS programs has done without bringing those measures forward to NQF.

DR. ROMANO: Yes, I think it is an important question. I mean ultimately at the end of the day it is a choice of CMS or others that are in this space as far as whether to
bring their composite to NQF for endorsement. So CMS, others are free to construct a composite and not bring it NQF for endorsement. By bringing it to NQF for endorsement, I think that there is an implication that they are prepared to defend it, that they are prepared to say that it is based on a concept that they can defend, not just because the individual measures are useful but because the overall measure, the composite measure has properties that make it useful to decision makers.

So what I am hearing in general is the sense that what we want going forward is for measure developers to present a very clear, articulation of what their measurement concept is and how they designed their composite to operationalize that concept and related to that, how they intend people to use it, how they intend it to inform the audiences, decision-making or whatever.

And this is honestly this where I
am kind of with Liz in terms of saying well
isn't it all about risk. So this argument
maybe should be in steering committees. You
know you, others would have to defend your
concept and say well we think this is a useful
concept. And others would say well why is
that a useful concept? Because it all comes
down to patient outcomes. And if we are not
about reducing risk and improving outcomes,
why are we doing this at all? And then you
would come back with a counter argument. So
that argument can take place in steering
committees but probably we can't forestall
those arguments here.

I think that is what I am hearing
is we just want clear articulation of these
issues. Is that right?

DR. CHASE: So I just want to test
that because I would agree. But when a non-
composite measure comes forward are we putting
a new standard on there beyond the composite
itself? Because it sounded like you were
saying you have got to bring your theory of
how it would be used, which I would agree
with, but do you have to do the same thing if
you are bringing forward a single dimension
measure? And I think the answer is yes.

DR. BURSTIN: Yes.

DR. CHASE: And so all we are
adding in this is saying you have to do the
same thing in around why it is being
composited, as opposed to just again the
individual components that you can bring.

DR. ROMANO: And how it is being
composited because it may influence the
weighting methods that you use.

DR. SHAHIAN: But to Helen's
hypothetical 50 measures that we use and I
would just like to roll them up into one, is
it sufficient simply for one to articulate
that vision of their composite or do we
challenge the empirical basis of doing that,
as Sherrie has suggested?

I am not sure it is sufficient in
my mind simply to have the developer
articulate that that was what their vision
was. Otherwise, this becomes a free for all.

DR. DE LONG: I absolutely agree.

I am so much less interested in how they
developed it as how it works. And it seems to
me there has to be evidence that it is doing
what it was intended to do. And if you just
mash together a bunch of outcomes or processes
and they are actually shown to be effective,
it doesn’t really matter what your perspective
was going into this, I think.

DR. SHAHIAN: And vice-versa.

DR. DE LONG: But you have to have
good evidence that they are valid. And we now
have some benchmarks to test against because
we have different types of measures.

For example if you bring forth
something developed using item response theory
and whatever and you test it against itself in
a split sample, you test it against some of
the other types of weighting so to speak, all-
or-none or whatever and it holds up, you have
brought forth evidence, it doesn't really
matter that you used item response theory
versus hierarchical modeling.

DR. KAPLAN: I think you need both
things and here is why. I think you need both
perspectives and the reason is suppose I had
the two trains left Chicago and a bunch of
questions like that. And so they were all
consistent and everything was great, except I
wrote them in French. And so what I really
was measuring was your ability to understand
French.

So yes, you can create composites
out of stuff but you will end up with Helen's
problem of adding up apples and airplanes
unless you have some expert-defined conceptual
approach that this is measuring X. And I
think it is measuring X. And then you have to
test it and make sure it meets the measurement
science standards of performance but you
bloody well better have a first idea that you
weren't measuring French instead of algebra.

So you have got to have that
undercurrent of understanding conceptual
grounding. And I learned this the hard way in
clinical circles, you had better be able to
defend that this measures diabetes care and
not patient sloth or something else that is
undercurrent. Maybe I attracted a bunch of
patients who are really, really couch potatoes
and lazy and yes, okay, it was my job to get
them to exercise but I can't come home and be
their personal trainer.

So you have to have measures of
what you are actually trying to measure,
physician performance, hospital performance,
whatever, grounded in some kind of conceptual
base.

DR. ROMANO: So I think in
response to David's, I think the solution here
is not lowering the bar. I'm not saying that
a developer can say anything but saying
basically that the developer has to articulate
the measurement concept and then show empirically how what they have done is consistent with that measurement concept.

And so that implies then the Steering Committee can have two separate discussions; one, if they like the measurement concept, one if they accept the empirical evidence, which might be different.

And so if I have been in the meeting with Jim, Liz and I would have argued against his conceptual framework but we might have perfectly accepted the empirical evidence but fundamentally those are two separate discussions. Is that fair? But both of them need to happen.

DR. ZASLAVSKY: I suggest that maybe the action item here is to commission the paper that organizes it in a kind of simple way some of these different kinds of rationales and conceptual bases for and give some examples, as Sherrie says so that people -- is it not going to be a dropdown menu. Not
in our lifetimes and hopefully never. But there are a number of useful concepts that are rationales for putting things together in particular ways and doing certain kinds of analysis. And people can be encouraged to use one or maybe more than one would be better because if something can be justified on more than one basis or evaluated on more than one basis, you have a strong evidence base and this would be something that people submitting these would be asked to refer to in developing their own short statement of conceptual basis and intended use of their measure.

DR. ROMANO: I like that idea. I mean there is some previous work that we can draw on, both I know Sherrie wrote a previous paper I think for NQF. You have written some work. So I think there are some elements that potentially we could borrow from. And obviously NQF staff will have to make a decision about the commissioning per se.
that recommendation?

MS. PAGET: I like that idea as well but I guess I think this is the place for this comment and maybe it is editorial and maybe it is more than that. But somehow it would be really advantageous if we could, and I think this is in our jurisdiction also help define for these measure developers, when you use terminology such as optimal and your data source doesn't include the patient, I don't know if it is optimal or not. So optimal or universal or any of this kind of terminology that implies that boy you hit that composite and you are golden and yet no one is deriving systematically any information from the patient him or herself. Somehow I would like to be able to embed that into some of these principles that we are talking about.

DR. ROMANO: Yes, okay so let's put that on parking lot. There may be more terms that we want definitions around. So when people use the term optimal or when
people use the term -- there are other terms
that we have seen. Maybe we ought to have
more definitions so let's put that on the
parking lot.

Before we take a break -- oh, I'm
sorry. Was there another?

Oh, before we take a break I would
like to see if we can get some discussion or
some agreement about these other conceptual
models that Karen and her team has summarized
here. And I think that the concept here is
that although we don't necessarily have to
have specific bins within composites, we do
have to have a process where people declare
whether they have a composite or not. So that
we can't avoid. So because NQF has to have a
separate forum. They have to have a separate
process for evaluating composites.

So there is a need, I think, to be
clear about what actually represents a
composite and what doesn't. And as Karen has
suggested, there has been some inconsistently
in previous NQF processes about that.

So let's see if we could -- so the specific terminology that NQF has used is that a composite is a combination of two or more individual performance measures in a single measure that results in a single score. So going to table three for example, one implication of this we think is that when you have a "composite" that is actually conceived of as a single measure that is based on multiple items, that would not be considered a composite from the standpoint of NQF review. In other words, that is just saying that in order to assess this concept, you need to ask seven questions, seven items. And each of those items is not a performance measure in itself. Those are just the components of a single measure. A measure of communication, a measure of timeliness, whatever it is.

So that would, I think exclude many of the CAHPS-based composites, wouldn't it, that have actually been reviewed as
composites?

So what do you people think about that? Do people see that? Because again the idea here is that NQF is treating a composite in a little different way maybe than what psychometricians do which is that a composite is a composite of measures, not a composite of survey items.

DR. BIRKMEYER: Can I see if I understand the distinction that you are trying to draw? Because I am not sure that I do.

Are you saying that an instrument that has multiple items that are trying to measure one thing at the patient level, that we know is not a composite measure. But things that are rolling up measures of performance at the provider level, those would be.

So is this just a patient level versus higher level distinction that you are trying to make or is it subtler than that?

DR. ROMANO: No. No, no. I don't
think it is subtler. It is just different.

I think if you have, for example, and again I am going to defer to the CAHPS experts, but if you have a number of items that are necessary to create a reliable measure of physician-patient communication, that is a single measure of physician-patient communication that is built on a set of items. But if you then say that I am going to create a composite measure of patient experience with physicians or patient experience with hospitals that rolls up five different aspects of the patient's experience with the hospital, that becomes a composite that NQF reviews differently as a composite because it is built on five different measurements.

DR. BIRKMEYER: This is a question are the measures rolling up to one domain of performance or multiple domains?

DR. DE LONG: That's not how I see it, actually. I see it as you could have an instrument that is for the purposes of
assessing patient-doctor communication. You could also apply a question to a doctor. That is the same domain. How do you communicate with your patients? How much time do you spend with them and whatever? That is in the same domain. It is measuring the same thing but it is not -- it is a different item, a different measure.

The other one was a survey that you asked and it had 36 items and the patient filled out all 36, we hope. That is an instrument that has its own measurement properties. You could roll up those two as a composite. They are in the same domain but they are not the same thing. But we are considering all the questions in the survey as in one bucket as the outcome of that survey is the measure.

DR. ROMANO: I mean just -- oh, Helen wants to speak. But just to give a specific example.

So in measuring blood pressure for
example we commonly say that people need to take two or three blood pressures measurements and average those. So conceptually, that is the same thing as asking two or three items on a particular domain and saying that we have to use that in order to construct a reliable measure of a single clinical concept.

DR. BURSTIN: Well let me just follow on the CAHPS example because that is actually one we have thought a lot about and Liz and Alan and others can weigh in here.

So for example we don't endorse the CAHPS survey. Very clear. Lots of items in the individual CAHPS survey. We don't endorse the CAHPS survey or the items within it. We do endorse the score based -- a performance measure based on the use of the CAHPS instrument.

The question is, if the CAHPS performance measure includes five domains that are separately reported out and publicly reported as separate domains, is that really
any different than individual measures, if
each of those components becomes essentially
a measure to be publicly reported? Is that in
some ways a composite performance measure?

DR. ROMANO: That is what we are
saying. Under this framework, this would be
viewed as an individual measure and not
treated as a composite measure.

DR. BURSTIN: The CAHPS? A CAHPS
performance measure would be an individual
measure rather than a composite, even if it
has individual components to be publicly
reported?

DR. ROMANO: Oh, no. It is when
you are rolling up the five domains that it
becomes a composite. The individual domains
are not composites.

DR. BURSTIN: Yes.

DR. KAPLAN: Okay, now I am really
lost.

So back to sort of where I am
trying to kind of get my arms around this.
There are things we call higher order constructs like mass that each have -- I am trying to kind of -- Shelly hates it when I do the math example, but my husband is a physician so he likes sort of more clinical examples. But math is a good one because it has algebra. It has all of these components we are all familiar with. But if I used a single item to estimate any one of those individual things, you would be very unhappy. And if I published that single item like algebra with one question, you would be really unhappy.

Those are also complex constructs. They need multiple items. So just like CAHPS has patients' experience with doctor communication, did you like the front office? The facilities, were they clean or dirty, blah, blah, blah.

Now I am going to create a higher order construct and wind those all up into a score.
So we have got these complex constructs and then we have higher order constructs. Supposing you wanted to measure obesity. I can use, by the way, also a composite, the body mass index. Then I can add truncal obesity. Then I can add whatever water displacement super-duper thing I have at my hands. Then I can add something else. Your reports about whether or not you gained a belt size in the last six months or a dress size in the last six months.

I can add all those things up and I am creating a higher order construct called obesity. So I think we are getting caught up in this composite business. Anything that you use more than one thing to estimate, in my view, is a composite. And what you do with that afterwards and what you are trying to represent become the construct you are actually trying to represent. And as that moves further and further away from and gets larger and larger and much more
multidimensional, that is when you get these
higher order things that have all of this
interior that could be separately reported but
also needs to be evaluated to the extent it
represents this higher order construct.

DR. ZASLAVSKY: I'm not sure there
is an entirely principle of answers to that,
the question that you asked Patrick in that if
someone came in with an SF-12 and said that we
wanted to report this out, you are probably
not going to start over and form a committee
to examine whether that is an adequate
construct, even though it is a composite.

The CAHPS items, there are certain
groups of CAHPS items that have been used for
12 years. You are probably not going to spend
a lot of time on that but if someone came in
with another version of the CAHPS survey, of
which we are working on about five of them
right now, and had another set of items that
were put together as a single construct, you
probably would look at that.
So it is really, I think, more a question of the history and the existing evidence base of former scrutiny, rather than any real difference in principle between what you are looking at in those different situations.

DR. ROMANO: Yes, I think what -- I am really trying to help NQF here. And so from the NQF perspective, NQF is about performance measures, not items, measures that are used to say something about provider or plan performance that are used to drive the market, that are used for public reporting, that are used for accountability and so forth. And so you may call those things composites just because they have five items that are all algebra questions. And of course it is obvious that you can't say anything about people's ability to do algebra without asking them at least five questions. You may call that a composite but from the NQF perspective, that is not a composite. That is
just that you need five questions to address this single concept.

What makes it a composite from NQF's perspective is that you are taking multiple measures that are performance measures that are separately reported as performance measures that say something about different domains of performance and you are rolling them up into a higher order, if you will, a higher order composite that addresses a larger concept. Is that helpful?

DR. DE LONG: So can we have clarification? You indicated that the SF-12 in your terminology is a composite. My understanding, according to what you are saying, is that it is a measure that has 12 items.

DR. ZASLAVSKY: Well let's take two really clear cut examples. If you do your three blood pressure reading was your blood pressure while doing jumping jacks, your blood pressure while lying on the table and the
blood pressure while eating lunch, those would
be three different measures that you composite
in a particular way.

If it is just taking three
randomly chosen algebra questions or three
measures just at randomly chosen times under
the same circumstances, then that is not a
composite. That is just replication of the
same measurement.

But if you look at the CAHPS, like
the CAHPS getting care quickly scale, asks
about getting care quickly when you need an
urgent visit and getting care quickly for a
routine visit, you know, getting appointments
for a routine visit. Those are different
things. The decision to put those things
together is based on a conceptual model of
their content, mainly of their content-
relatedness, although to some extent, based on
psychometric evaluation as well.

And so somewhere someone had to
look at that and make a judgment about that.
Once that is done, you are not going to revisit that every time you use that measure and you may, for future purposes, think of that as being one measure when you think about going up to a higher level super composite. But there was a process initially of treating that as a composite because it wasn't really three different items. The SF-12, you know, I don't know there might be -- whether you are sad and blue or happy and pink, or whatever the different items there, and they are different questions. And there was again some decision made about how you form those together, which we don't revisit every time we use it. But there was some kind of a process, a lot of process that people went through in order to get there.

DR. BURSTIN: Just one thing, and I am not sure if it is helpful but this was a major issue for us as part of this PRO work we just did. And I can't tell you how long it took for the committee to agree on what these
things are all called but let me just try this because I think it was helpful.

So in the PRO context, the patient reported outcome was the concept and we actually used the Minnesota measure of depression as a way to sort of explain this.

So PRO content is want to look at depression. That is the concept, the PRO. We then talked about the PROM. People talk about patient-reported outcome measures. In that case, it is the PHQ-9, which is a standardized tool used to assess depression. To me, the CAHPS is a standardized tool used to assess patient experience of care.

We then tried to make a distinction of a PRO-based performance measure, a PRO-PM as we called it, which was the performance measure based on use of the tool.

I think what Patrick is trying to say is we are not going to get into the issues of the tool itself or the components of the
tool. We will get into it only insofar as it relates to the use of the measure for performance assessment. And I think that is probably enough. And I don't think we need to do too much more on this.

DR. DE LONG: But one of our mandates here is to define a composite. And I think we are still not there with respect to whether the CAHPS survey is --

DR. BURSTIN: Wait a minute. Maybe I am off but I feel like we are there.

DR. DE LONG: Okay.

DR. BURSTIN: I feel like we have a reasonably good sense that the tool/survey is not the performance measure. It is not what NQF endorses. It is the substrate around which people develop a performance measure. We are only endorsing the performance measure. So there may be complex concepts as part of surveys. That is all well and good but when it gets to NQF, we are talking about the measure around it, rather than the survey or
the tool.

DR. ROMANO: So it is important, I think, that we need to be clear throughout that we are talking about composite performance measures, not composite measures -

DR. BURSTIN: Correct. Performance measures.

DR. ROMANO: -- but composite performance measures.

DR. BURSTIN: Correct. Yes.

DR. CHASE: So I would agree with this. I think we can be there with once there is a standardized tool that is not a composite. Because this is helpful because when you are submitting something you want to know if you have to check that box or not.

The one I think there were there is another area of gray that you mentioned in this, as I recall, is what about multiple sort of measures of a particular thing? And I will give the example of -- and let's not get into
the clinical stuff because I won't get that.

If we were saying LDL control, you could construct a measure that says you either your LDL was either below a certain level or you were on a statin. Those are two things and they are being combined. To me, that is not a composite measure. That is two ways to ask the same, to evaluate the same thing and that might be helpful, too with some guidance of when it is multiple pieces identifying something, unless you are somehow constructing it in a different way. When it is just multiple yes/no to get to the same question, that is not a composite.

DR. ROMANO: So you are talking about sort of Boolean logic in general, where there is a set of and statements or or statements that are necessary --

DR. CHASE: Yes.

DR. ROMANO: -- for the construction of the measure.

DR. CHASE: Right. Again, you can
take it to a far degree where we might say now
you really are a composite. Because again I
think these can morph into some gray areas.
But there are a lot of things that come
probably to measurement where there are
multiple things being assessed but they are
really still the same thing.

DR. BURSTIN: And this comes up
with us a lot, people submitting measures as
measure pairs. Always look at this measure
with this measure. And we struggled actually
about whether or not we should bring to you
the issues of pairing, tripling, and
quadrupling and we decided not to for your
sake and ours because it is a complex issue
but we don't believe those are composites
either.

DR. BIRKMEYER: Well then I don't
get the definition, then. And I appreciate
like how simple that example is. You know,
you have got a process measure that a person
is or isn't on a statin and you have got some
continuous measure of the LDL and you are combining that around a construct of LDL management.

And I don't understand -- and they are measuring different things but it is under the umbrella of a single construct and I don't understand how given what the measure is of a composite, why that is not a composite.

DR. BURSTIN: Because I think they are not separate measures. I mean let me play that out because actually that is a good example.

We will sometimes see measures come forward and again, depending on the data source, they may say you can capture LDL control in one of several ways. One way is to actually be able to have the actual laboratory data and say LDL is less than 100.

One other way may be and there is pretty good evidence in the cardiovascular to say statin alone is probably good enough.

So in some ways even irrespective
of the LDL level that you would actually combine those constructs and say either of those meets the numerator for this measure. It is not as if they are two separate scores, two separate measures combined into a single score. There are different ways of representing, I think the same concept. It is fuzzy, John. I’m with you.

DR. ROMANO: But I think that it is the same. It is the same thing that we just talked about, which is that in order to measure a single concept you have to look at two different pieces of information. Another example is very common in clinical studies to assume that anybody who is on an antidepressant as depression, even if their PHQ-9 score is fine. So it is the same thing. If you define depression as being on an antidepressant or having a PHQ-9 score, you have predefined that that is what is necessary to define that concept. It is not compositing two different performance measures. It is
saying that you need two different items of
information to address a single performance
concept.

DR. BIRKMEYER: I'm sure that I
will learn more but I still haven't quite
gotten to the point where this is anything
more than just a simpler version of the
advanced diabetes care instrument roll-up and
a much simpler version of the STS version
where you are taking like one measure of a
process of care and another version of an
outcome.

So but I will stop talking.

DR. ROMANO: Well I mean I think
we might or might not agree with that
particular example. I mean, I think Helen's
example may be clearer. So you may push back
on Jim's example and way well that is not a
good example of the phenomena.

DR. BIRKMEYER: Well I certainly
get kind of the multiple items within one
instrument that gets applied at the patient
level thing as like one measure and then gets
rolled up to a provider. That I get. And
that is what the opening document of NQF says
and I get that distinction. But this other
example to me feels like it is very different.

DR. KAPLAN: I'm worried and I
don't want to be responsible for delaying our
break here but I am worried that this is --
I'm lost, too. Because for me the data source
is irrelevant. If your survey -- I don't
care. If it is a multi-dimensional construct
and it comes from a survey, so what? It is
still a multi-dimensional construct. What I
thought Helen was originally saying was we are
treating those little multi-measure and don't
get lost on the item versus measure, every
single one of these things is a measure, it is
just collectively they measure a different
concept or a concept together.

But the problem that I am having
is you are really, I think, NQF is talking
about higher order constructs. For me, you
are starting to add up patient experience data
with hospital mortality data, with patient
safety data and now you have got a real mega-
construct about how good this hospital is. Do
I want to go there? Whatever that is, that is
a real mega-construct.

Now you have got higher order
constructs that are combining information from
various different sources and that we
shouldn't get lost on. They could all come
from the same data source but they measure
different things that collectively now measure
something larger.

And so if that is NQF's definition
of a composite, then strike the one that is in
the document now because that is confusing.
It confused me. Just adding up two or more
things, it depends on what you are trying to
represent.

And I think, Helen, what I
understand you guys trying to represent is
something larger than the patient experience
data or collectively or even taking all the
patient experience data collectively. It is
something higher order than these measures of
stuff that can be multi-item, multi-component.
But it is a real higher order construct that
you are talking about.

DR. BURSTIN: It is always higher
order. I think at times we have seen -- I
mean in the last year we had a cardiovascular
project. A series of measures came in for
patients who had implantable defibrillators.
They should be on this. They should be on
this. They should be on this. Cardiovascular
was like, this is nonsense. They should be on
all of those. Make it a composite. So ACC
took it back and made it a composite.

It is not necessarily something
higher -- again, I don't want to get hung up
on what is higher order to me versus higher
order to you. I just think the end of the day
the idea was those individual measures they
stood alone told very incomplete parts of the
picture and the cardiovascular committee was left uncomfortable that anybody would use any one of those measures in isolation and assess the quality of care that that cardiological service was providing for patients with ICDs.

DR. KAPLAN: Well let me come back to them and ask a question. Supposing I have a new measure of participatory decision-making so it is not a new measure. But supposing now I have -- I am trying to get a sense of whether or not doctors include patients in treatment decisions. And I have seven items that measure that. They all come from a survey. That is a composite measure but --

DR. BURSTIN: A composite performance measure.

DR. KAPLAN: Okay, now I am still stuck on -- and if I am lost, the odds are at least 50-50 that somebody else will be lost.

DR. BURSTIN: I think it is measure/tool -- I mean, people call those things all kinds of things.
DR. KAPLAN: Well first of all I like instrument because a tool is to dig and shovel and an instrument is to make smaller things with. But if you have got -- so then you have to, I think, inform the field about what you are meaning by performance assessment composites and make the definition more related to that. Because it is still performance if the physician is being evaluated on a set of skills if they come from the patient, they come from the chart, they come from here. They are still being evaluated on a set of skills. If you don't mean to include interpersonal care as that set of skills, then that is an important distinction to make.

DR. ROMANO: Well we are overdue for a break. So I think we might have a couple of offline conversations during the break but everybody rejuvenate themselves on coffee or whatever your preferred beverage is and we will come back in ten minutes, I guess.
(Whereupon, the above-entitled matter went off the record at 10:55 a.m. and went back on the record at 11:13 a.m.)

DR. ROMANO: Well let's go ahead and reconvene. Dale, are you still with us on the phone?

DR. BRATZLER: Yes, I am.

DR. ROMANO: Wow. Thank you for your patience.

DR. BRATZLER: I set aside the whole day to be completely available.

DR. ROMANO: Okay. We can't read your body language so just feel free to interrupt as you deem appropriate and we will respect that.

DR. BRATZLER: Yes. It was a very fascinating discussion. I'm not sure I could have added much.

DR. BURSTIN: Dale, this is Helen. So I think we are going to talk a little bit about all-or-none. You might want to give us some of your insights from those measures you
guys were using in the QIOs as well.

DR. BRATZLER: Okay. So would you like for me to go ahead at this point?

DR. BURSTIN: Not quite yet.

DR. ROMANO: Wait a second.

DR. BRATZLER: Yes, okay. All right.

DR. ROMANO: So I think that the bad news is that we have gotten through one out of a whole list of questions.

(Laughter.)

DR. ROMANO: But the good news is that that question is so big it has really encompassed some of the other questions within it. So we may have made more progress than we think and we have certainly come to recognize the complexity of this space.

I think one of the lessons that came out of my offline discussions during the break is that really the measurement science tools that many members of this committee bring to the enterprise that these tools need
to be brought into the discussion of all measures, including measures that we may describe as not being composite measures for the purpose of this discussion. And this kind of leads into some discussion of these Boolean measures. And so I wanted to get those issues on the table and then we will go into a little bit more about this business of component measures.

So we have had a number of measures that have had Boolean logic, either a series of, if you will, as Alan mentioned in the break, serious reportable events where it is a series of things that did the patient have this, or this, or this, or this, A, or B, or C, or D. It is a long list of complications, typically.

Those are sometimes viewed as composite measures. Sometimes they are viewed simply as a single measure that may have several different components.

Similarly, we have these all-or-
none composites that are based on the premise
that providers must do A and B and C and D and
E  If they do all those things, they get
credit. If they don't do all those things, they don't get credit.

And so the question is are these
scoring methods for composites or are these
different types of measures? Should these be
viewed -- and I think our discussion, your
chair's discussion with staff before suggested
that these types of measures should really be
viewed as single measures, where the
developers are coming to NQF and saying that
we think that in order to measure this
construct, it needs to be done with a series
of ten questions and it needs to be formulated
as A and B and C and D and E. And so the
notion then is that these are not separate
performance measures but these are ten items
that are necessary in order to tally a single
performance measure.

So is that a distinction without a
difference? Is that useful?

So with that construct then, all-or-none scoring and any from a list would not be considered composites for a separate review process. They would go through the process being considered as individual measures. Is that -- am I summarizing that, Karen? So what do people think?

DR. BRATZLER: Patrick?

DR. BIRKMEYER: I'm sorry. We discussed this a little bit at the break. It seems like we are trying to draw a dotted line at like what altitude do you need to get to be a composite lover. But at the end of the day, does this matter only to the extent of like which committee or group to these measures go to or is it more important than that?

DR. BURSTIN: It is more substantive in that it then leads to a whole set of questions about whether we need to get into a deep dive on the components within the composites.
So this is one issue. If an all-or-none is not considered a composite, then how do we handle the components within them and the efforts to harmonize with existing measures?

DR. BIRKMEYER: But just to follow-up, you know, if it went to the composite evaluation process versus the regular, would there be a different level of scrutiny on the components that roll up into whatever it is that is being measured to get more slack of you go one way or the other?

DR. BURSTIN: According to our criteria, yes. In that --

DR. BIRKMEYER: Which is more stringent?

DR. BURSTIN: The composite measure evaluation criteria would specifically require that the component measures either be evaluated to see if they are stand-alone measures or at least meet criteria for appropriateness within the composite, even if
not endorsed as stand-alones.

DR. BIRKMEYER: So if you had a measure that could be demonstrative -- whose validity and usefulness could be demonstrated at the summary level but not at the component level, then it would be one would preferentially not want to go through the composite evaluation process. Am I understanding that right?

DR. BURSTIN: I actually never thought of it that way before but I guess that is one way to flip that on its head. I always think of it in the other direction of saying then you have to go look at the components as opposed to the flip of not.

DR. BIRKMEYER: Because I would have guessed just the opposite. I mean, before I got into this process I would have guessed just the opposite, that the whole process of breaking out in evaluation process for the composites is the focus on sort of the measure characteristics at the summary level.
and not at all of the little pieces.

We already have a process in place that can evaluate all of the little pieces.

DR. BURSTIN: Should the little pieces then be submitted separately?

DR. KAPLAN: I wasn't going to say anything before lunch --

(Laughter.)

DR. KAPLAN: -- but now I am a little bit -- I am even more lost. Because if -- take the diabetes -- I hate to harp on diabetes but it is an example I know the best. If you were going to say and we just told the National Association of Public Hospitals to push back CMS on this very issue about the all-or-none scoring for the diabetes measures taken as a group. So if there are nine of them, say, and I was going to create a composite for institutional-level performance and it was going to be an all-or-none, it wouldn't go to the composite. You have to have A and B and C and all the way to nine.
It wouldn't go through the composite process but if I was going to say I am going to evaluate how many of these you got and give you a score however I do that, then it would go through the composite process.

DR. BURSTIN: That is why we have always treated all-or-nones or weighted composites the same. And they do go through the composite process.

DR. KAPLAN: All-or-none scoring of it would still not be enough -- would be enough -- wouldn't trip it into some separate review process. It is still a composite?

DR. BURSTIN: In our current parlance, it is still a composite and that is a question for you. Is that reasonable?

But at the same time we do ask the committees to go through -- we just went through this. We have a colonoscopy quality index that just came to NQF, was submitted on a single form, all nine components or ten components on a single form. The committee
had some concerns with two or three of the components out of nine of being perhaps not as evidence-based as they would prefer.

So at the end of the day, do they thrown out the entire composite because two or three of them they didn't think rose to the level of the others? Should that have been submitted on nine separate forms so that we actually can take the deeper dive. Ultimately we will end up re-reviewing the measure and force the committee to go component by component because otherwise we just couldn't make sense of it.

DR. ROMANO: Well, okay. So let me get radical here, which is so I am feeling a lot of confusion and a lot of push back or concern about sort of trying to draw this bright line between what is a composite and what is not a composite and what undergoes composite review and what doesn't.

So maybe we should throw out this whole distinction and just go back to measures
and we just have measures. And so if a
measure developer comes to NQF and says I have
a measure. It happens to roll up five other
measures but at the end of the day, it is
supposed to measure some concept. And that
measure can be evaluated using the appropriate
tools.

Is it conceptually -- I mean I
think this is getting to your point. Is it
conceptually different to have all or none
scoring versus some kind of weighted scoring,
to have a separate process for those two to be
evaluated? An easier process for one than the
other doesn't feel right. Is that what you
are getting at? It doesn't feel right.

DR. KAPLAN: Yes. To me, you have
got nine things that you are measuring and
your collectivizing them some way or the
other, the scoring is irrelevant. You are
still collectivizing them. If you turn them
and score them, if you all-or-none score them,
if you do mean scores, you are still pulling
together a set of things that measure what we are going to call diabetes quality. And maybe I don't even have enough of them for certain levels of performance assessment or maybe I can get away with fewer at other levels of performance assessment but together they measure diabetes quality. And that is what, for me, represents a composite.

DR. CHASE: So I would agree. I don't think it makes sense to just totally -- so all-or-nothing composites never have to come through a process. But how they are dealt with once they are there could clearly be different because I think what would be nice to avoid is -- and correct me if I am wrong. But it has felt like sometimes there is a discussion that has gone on about those that like all-or-nothings and those that don't. And so that becomes the discussion, as opposed to if an all-or-nothing measure came through and its four components are all endorsed measures, what seems to be in front
of the committee is really just a question of

do we think those four things hang together?

Do they make sense? There shouldn't be a lot

of additional review about it, whether an all-
or-nothing is the right to do, especially if

NQF is saying as a principle, we endorse that

under certain circumstances.

There are other situations where

if somebody is bringing a composite that

weights those in different ways, then I think

there is an extra step, which is does the

weighting make sense. I mean then I would

think that you move into a different direction

with the committee of saying you also need to

assess whether the weighting is okay.

Now that may also not be a fair

distinction because why shouldn't the

committee say we want to assess whether all-
or-nothing is a kind of weighting. And so we

should have the same rights to do that.

So maybe there is no distinction

there at all.
DR. BIRKMEYER: But I actually get that distinction. Kind of the all-or-none measures for which at the end of the day there is no empirical criterion standard at which you are assessing this composite against. There is no science underlying the weighting, other than the collective judgment of somebody that put these things together.

In that particular instance, sort of the validity of the components is really the only thing that you could assess. There is the complete opposite end of the spectrum with composites that they are being evaluated against some criterion standard like mortality with a procedure and all of the science is not around the components and whether they are valid but the statistical or other weighting approach by which they get put together and those to me feel like they need to be evaluated with a different lens. And I don't care like where they go or what they are called, but they are very different.
DR. ZASLAVSKY: I will just repeat the point I made to Patrick on the break is that the kinds of issues that come up when you put together, call them measures or items -- I guess we are not calling them measures if they are not reported out. Whatever we are putting together, the kinds of issues that come up are some of the same issues come up regardless of whether you are putting together a bunch of things that are never events in surgery or something like that or putting together items on the CAHPS scale. And there are different ways of doing that. And what I am concerned about is that there should be people in the room when those measures, those consolidated measures, whether you call them composites or not, are being evaluated who understand some of the issues in deciding to do all-or-none scoring versus a weighted scoring versus an equally weighted scoring, you know, different options that you might do and that that kind of thinking is part of the
evaluation.

Whether administratively you want
to track things as being composites or not
composites, you know, that is more of an
internal NQF matter which I don't have an
opinion but it is the kind of thinking that is
brought to bear on analyzing these things.
Because I don't want a whole bunch of things
to be done as all-or-none things because
surgeons think of it that way or because
endocrinologists or cardiologists or whoever
is involved thinks about it that way without
having also some of the statistical
measurement expertise brought to bear in
examining that. And I am looking at whether
that makes that sense as the best of reporting
out that information.

DR. DE LONG: So I am under the
impression that not only is there a different
process but developers have to declare whether
it is a composite. And they have to follow
certain guidelines that are specific to
composites. I sort of side with Patrick that maybe trying to make a distinction and forcing them to recognize some of these things that we don't seem to be in tune on could create more confusion than is necessary. But I am not sure what hoops they are jumping through that they wouldn't ordinarily have to jump through if it weren't a composite.

DR. BURSTIN: I think it is really just what we are going to -- and you have it in your packet, the distinction of what are the additional requirements around evaluation and submission that is different? And much of this comes to the construct, how it comes together, the testing around it. And at least I think an important distinction from where I sit is we are spending so much of our energy these days focusing on related and competing measures is how do we handle the component measures within them? And do they need to get fully evaluated on their own in a way that allows us to do the related and competing and
make assessments of whether or not they are in fact best in class. It becomes very complex out there, whether there are measures that are living in composites that don't necessarily relate to the measures you are being paid on for other kinds of purposes. So how do we make sense of that?

So for me, I am being less theoretical and more just practical of what do we ask developers to submit? What do we ask committees to consider? And those criteria will be important.

DR. ROMANO: So initially NQF created a composite measure evaluation committee that produced this report that we all have. But going forward, perhaps you could elaborate a little bit, my understanding is that NQF is expecting the individual topic-specific steering committees to review composites within their clinical or subject area domains. Do we have people on each of those committees that have the kind of expertise
that Alan is describing? We won't call it psychometric expertise. We will call it measurement science expertise. Do we have people on all the committees that understand these concepts about that the pros and cons, the strengths and weaknesses of different ways of combining multiple items or measures together?

DR. BURSTIN: Right. We strive to at least put a couple of methodologists on every committee but it is a couple. It is not like this room, certainly, where the methodologists outweigh poor country internists, as I will put myself in. But you know at the end of the day we do have people sitting there.

But one of the approaches we have taken which we did as part of our large outcomes project a couple of years back, is every outcome measure got reviewed by a statistician. Actually, Sean, which was brilliant, worked really well. And one
question might be as a recommendation to this committee is that we say outcomes and composites should have a statistical review. Almost like the annals always sends papers to statisticians for a secondary review for complex models. Is this complex enough that you think that the average committee member couldn't necessarily handle it?

I will tell you we have seen the committees not necessarily get as worried about the constructs as much as the fact that this measure has all the components I think are really clinically important. It is often more clinically driven I think than it is methodologically driven.

DR. ROMANO: Yes, and at the end of the day this whole discussion is about how to make NQF's processes work better to provider clearer guidance for developers and for steering committees.

DR. KAPLAN: I think that the world has changed, as they said in one of
those trilogy of "The Lord of the Rings,"

because I think CMS amped up the stakes up

when now they are going to start -- you know,

when these are starting to be used for

compensation and I think one of the committee

meetings next week is going to look at some of

those very issues.

But I think the stakes are

different now and I think NQF is right to

worry about when you create these measures and

they get put out there and they get used for

paying people, you just ratcheted the fire up

substantially. So I think the shakedown might

even -- I don't know, Helen, what you think

but I think the shakedown now might have to

look different as things go forward.

DR. BURSTIN: Part of it Karen

just reminded me as well, as part of our

submission form, we already have a whole

section on risk adjustment if it is an

outcome. And maybe we don't have a separate

composite form per se but that every single
submission form, if it is a composite, answer
the following set of questions but not
necessarily create a whole separate event, but
maybe have those risk adjustment and those
composite approaches considered by experts as
well.

DR. ROMANO: So I think that is
sort of where we are sort of going that
instead of having a completely separate
process, that this would be viewed, the
composites would be viewed within the
ordinary, if you will, measure endorsement
process but that there might be some specific
declarations that measure developers are asked
to make to articulate what the higher order
composite, what the construct is, if you will,
and then what -- how they formed the composite
based on that construct. Does that summarize
people's views? Is that --

DR. DE LONG: We're still talking
about development because I think it does go

further in terms of evaluation and experience
with the measure as time goes on.

DR. ROMANO: So from that framework then if I am getting this correct, then we are not going to worry so much about all of these different conceptual models, three, four, five, six, seven and eight because basically developers would simply be asked to explain what their model was, what their measurement construct was.

DR. BURSTIN: And then if you look through it, I mean the only reason to actually I think go through these a bit is there are some special considerations, for example, around the submission. So for example, what you require is listed under one of these specific testing at the performance score level for composite measures. So we currently allow testing at either the data element level or the performance score level. If it is a composite, does it have to also always be a performance score level? We will come to that when we start going through the criteria this
afternoon.

DR. SHAHIAN: So will -- if we were to combine this in one measure form, submission form and no longer had a separate composite, would we be in any way diluting the additional requirements that we imposed in our previous document for a composite measure or simply be putting them into a different pathway once you declare your composite?

DR. BURSTIN: I think it is the latter, yes. So we would just not have a whole separate forum. We would just try to build in whatever those components are, if this is a composite answer the phone and questions as well.

DR. ROMANO: That leads into, I think the third bullet point here, which is there is this guidance that NQF has had before that the components of composite measures have to be separately endorsed or have to meet the criteria for endorsement.

So does that even make sense? I
mean if we are taking this broader view of
composites, then it may be that for certain
types of composites there is no need to even
consider the performance properties of the
individual components. For others it may be
more important. I don't know. Do people have
thoughts?

DR. ZASLAVSKY: I think I want to
return to a comment I made earlier that it is
sort of a question of bias versus variance.
At least that is the simplest way to look at
it and if something would fail as an
individual measure because it is
insufficiently reliable, like something that
happens in one out of every hundred cases,
that doesn't exclude it from a composite
because you are putting it together with 20
other things that happen in one out of a
hundred cases and you are up to 20 percent and
there is a fair amount of information there.

So that is the kind of criterion
that you would not have to apply to the
individual components. But if something is not acceptable as a performance measure because it discriminates against hospitals that treat sicker patients or because it is something that should be adjusted for age and it isn't or something of that sort, then that criterion should be applied to the individual measures.

You might, in some cases say that I am putting together A and B and A catches this group and B catches that group and when you put them together you get something that is fair, that would be legitimate. And so that would be a case where you have kind of offsetting biases of the different components.

But in general if something really is either scientifically not valid or the data usually can't be collected, if you have any of those kinds of criteria or it is biased in the senses I have been talking about, then I think those criteria would apply even if it is being put into a composite.
DR. DE LONG: So for example with the HDL -- was it HDL or LDL -- you had an either/or. And I think what you are saying is that those two measures would not necessarily be endorsed as stand-alone measures because they don't really tell the story. You wouldn't want to grade somebody on having the LDL lower than something if the patient was on a statin. But some measures really do need introspection.

DR. ZASLAVSKY: Yes, that would be a case where either measured by itself wouldn't work and you have heterogeneous patient populations with regard to their risk for hyperlipidemia but put together you have a fair measure of what cuts to this appropriate practice.

MS. PAGET: How does this impact the unpacking ability requirement that currently sits in the criteria?

DR. ROMANO: What do you mean by unpacking?
DR. BURSTIN: I used that term.

MS. PAGET: My understanding is it is actually called decomposition, I mean deconstruction, something but decomposing is just like -- so when she said unpacking this morning I said oh, I like that a lot better. But my understanding is that we want to be able to do kind of like of a "Consumer Reports" type thing where you get the red circle but then you can also look across the chart because I often find that there is not a lot of difference sometimes. And even though you have got a bright red circle, it is only a difference between a 94 and a 96 or something to that effect.

So I think the unpackability is important and I just didn't know how this relates to this issue of the individual components being NQF-endorsed.

DR. DE LONG: It seems that the developer should need to specify how this measure if it contains more than one indicator
should be unpacked. Because in the example we
were just talking about, I don't think you
would unpack, would you?

DR. KAPLAN: To me composting is
a lot like composting. So if you put together
a bunch of stuff in the compost heap and you
end up with fertilizer. But if you take any
of the eggshells and other stuff in there out,
they are still eggshells. But once they
interact with all the other stuff, now it is
fertilizer.

So decomposing, if you will, these
measures in some cases may not make any sense.
But making universal out of it isn't going to
work for you either. Because if you are
trying to -- some of these things may very
well stand alone and you are creating this
higher order or whatever we are calling it,
composite, to represent something else.

So that is where you get into the
now you are going to have to put people
through a different level of -- it is the same
measurement principles but they have to
demonstrate that they are actually now
measuring that something else, whatever it is,
fertilizer.

MS. PAGET: I guess I would just
say I would still want to know the pieces and
maybe in this case, it is not meaningful as an
individual but knowing what went into the
fertilizer, helpful.

DR. CHASE: I was trying to
understand that practically, too, I think we
want to avoid -- we want to have some guidance
here so we are not making measure submitters
or committees rehash stuff that has already
been decided. So again, when a component has
already been endorsed, that should give you --
I don't want say it gives you a pass but it
certainly should be a different level of
scrutiny. Because that I have experienced in
the committee work is you get a different
committee together, they may make a different
decision with the same evidence. So you don't
like to see that happening but that is what
you will start to build if you make people
rehash the same question over and over again.
So I think that is good but I think there
should be a different standard then or
guidance is when there is pieces that aren't
endorsed and they are there for a reason, you
have got to justify that and talk about what
is the -- there is a different level obviously
of review for those pieces, without making
them be submitted on separate forms. I was a
little worried when you mentioned
deconstructing shouldn't mean you have to
submit each piece to different committees,
unless it is totally different.

DR. BURSTIN: So just to follow up
on that, again if you think about what is in
our criteria and much of what Alan was talking
about was really about the scientific
acceptability of the properties. A lot of
what we are often talking about is the
evidence, though, for the individual
components. And so the question is, you have
got to be able to look up, if we are saying
those are not currently NQF-endorsed measures,
they have not been reviewed, the committee
needs to have enough information to really
understand whether the evidence for the
components is there. I mean, I am agnostic as
to whether it is a separate form or some other
little box that pops up or something but you
have got to be able to provide enough --
assuming we want to continue on the idea that
assuming we think the individual components
are really important, they are not endorsed,
they may add huge value to people out there to
actually have those individual measures out
there, what do we need to do to get the
committees enough information to make that
assessment? That is sort of where I am stuck
a bit.

DR. ROMANO: Well does it -- if
the measure developer is supporting the
unpacking concept, the drill-down concept, if
you will, then I think the implication of that is that the unpacked measures have to be able to stand on their own.

But if the measure developer is not supporting an unpacking concept. In other words, if the measure developer is saying that you can't unpack this, that this is what is necessary to measure the construct, my construct, it might not be your construct or somebody else's construct, in order to measure my construct you have to include all of these items. Maybe some of them have been previously endorsed, some have not.

What is wrong with that?

DR. BURSTIN: I don't think there is anything wrong with it. I guess I was just questioning at times is the measure developer always as a single entity, not a multiple stakeholder environment, always the right group to make the decision of whether it should or should not be unpacked. If that is value to a consumer -- let's think about Lynne
for example.

If she wants to be able to go across the "Consumer Reports" page and in fact see which of the component measures are really driving that overall score, maybe one of them is really important to her not having a stroke.

I'm just making this up, David, a bit. But you know, if stroke prevention -- not having a stroke plus CABG is really important to her and she makes that decision, should it always be in the decision of the developer to say what should be unpacked and what shouldn't? Just playing devil's advocate.

DR. SHAHIAN: One of my concerns is as we give more and more flexibility to developers in how they navigate through the system and which path to take, I think there will always be a tendency to take the path of least resistance and to choose the path which requires the least oversight, the least
empirical justification and so forth.

So I am a little concerned that in
an era when we should, I think, be becoming
even more fastidious about how we approach
these measures, we may in fact be taking
things in an opposite direction and I would
hate to see that happen.

DR. CHASE: To make sure I
understand this, I think it would be helpful
to have guidance, too, around the opportunity.
You presented one earlier where I think it was
the GI where there were components that the
committee liked and some they didn't. It
seems a shame that the only alternative now is
to just say no. And if we say each submitter
should have the right to decide whether it is
packed or unpacked, people are always going to
say I think it all has to go together. But
the committee should have the ability to say,
we will endorse the four out of the five here,
we just can't go to five, and not force it to
have to go to a resubmission which, and
process-wise may mean you can't get it back in a door for a while because we are wasting an opportunity to get something valuable out there sooner.

So I think it would be nice to allow the composite committees to offer an endorsement of less than the whole if it makes sense to them. And the measure steward is going to get to decide whether they want to continue with it based on it. They could say nice for you guys but then you go collect the data, we are not going to do that.

DR. BURSTIN: The challenge there then is that the measure testing that has been put forward is based on what was submitted. And this was an issue that has just recently come up. If they would then say okay we hear you, let's take out two of those components, is their testing still valid? Is that okay?

DR. KAPLAN: Yes -- no. No. I mean if for example you have got a seven-item measure and now you are going to remove two of
those measures, you are going to take a reliability hit just because of the way we compute these composite or depending on the way you compute these composite measures, the odds are good that you are going to take a reliability hit. The more things I measure about a construct, the better I am about being able to repeat it. So the consistency is about repeatability, reproducibility. So more things are better.

If you take two things out, now am I good enough to compare one hospital to another? What is my measurement area going to be?

So without that kind of support, but I think you should be restricting people from adding things that they think improve the precision of an estimate is not a good idea. If we think that it is important to estimate a complex construct with more things, then what evidence does the measure steward provide that that does actually
contribute to the reliability of the estimate.

If on the other hand you are going
to just toss the whole thing out and not allow
them some rejoinder to say well, we are not
sure but here is the evidence and here is the
face validity for these new things and here is
our guess at what it is going to do, and let
them model it, how much it is going to improve
the precision of this estimate. Why make it
go back to ground zero and just toss the whole
thing out?

DR. BURSTIN: I'm not saying it
has to. I am just saying those are the
realities of what we try to do in the course
of a project. I'm just being very honest
here.

But maybe that does speak to the
question if you are allowing the developer to
make the decision of pack/unpacked, then maybe
you have to have some pretty clear
requirements on the statistical evidence you
put forward for your packing. And I think
that is where we have actually seen a fair amount of lacking in terms of saying if you actually pulled out these, what would you -- and frankly, we see very little of that when it is submitted to us, with a few exceptions.

DR. DE LONG: And I think that speaks to current tension between the proliferation of measures and having to record all of these data when some of them aren't really necessary to add into the mix.

So that proof that you mentioned, I think is very important that each component needs to provide some more information.

DR. BURSTIN: And not to forget about one of our other requirements is feasibility. There is a huge measurement burden associated sometimes with the collection of these data. And if you are adding components and they are not having a measurable impact on the outcome and it is a lot of work to collect them, they probably shouldn't be added. And then people also make
the case it is not always just data collection burden, it is actually opportunity costs of having to look at those 19 components to make sure you are doing okay on the composite.

DR. CHASE: The burden goes the other way, too, to the measure submitter that we don't want to make it such a barrier. Say you have got 19 components, you have got to tell us the validity of each one, when you only may have been able to test the 19 together.

And so that, I think there may be some of these where it can go one way or the other as far as that you could choose. In other words if the committee felt like all 19 aren't valid, then you might have to provide some data but you wouldn't always be required to do all 19 because the committee may accept it makes sense to have all the pieces together.

DR. ROMANO: Yes, so let's say one were to propose a composite that included a
bunch of process measures, for example. And let's say that each of those process measures or let's say that maybe some of them have been NQF-endorsed, some of them haven't but they are all just different processes of care. If the developer can empirically demonstrate that this composite is predictive of patient outcomes and that it identifies providers, hospitals, if you will, where patients will have better outcomes, then do they need to justify that there isn't bias in the measurement of each of those components? Or is it sufficient to say that the composite in itself has desirable reliability and validity properties and that supersedes, essentially, issues about the validity of individual components.

DR. BURSTIN: It is a great question, Patrick. I will give a -- and not to keep diabetes is such an obvious one, so I am going to keep beating on it for a second longer.
So let's say for example you have got great evidence you have submitted that the composite, the diabetes composite you have submitted is highly predictive of outcomes, clearly identifies patients where they would have better outcomes, but there is individual measures that directly compete with the measures inside the composite, such that clinicians are getting the individual measures also put forward to them and they are getting differing levels of performance requirements across those. Should there at least be a requirement that they are harmonized so that you don't wind up with a clinician, doc, whoever over here, saying I have got to do 140/80 over here but over here they have 140/90, they say A1c 8 base -- diabetes is of course the worst example because it creates more fights than any condition I have ever seen in my entire life with the possible exception of readmissions.

And so that is the issue for us.
Should we at least insist that there be --
even if you don't dive deep on the individual
reliability and validity of the components, do
you at least need to harmonize to the science,
I guess?

DR. DE LONG: Is there any reason
not to? Are there examples?

DR. BURSTIN: Yes, because people
have measures in use they have used for years.
It is not easy for people to flip on a dime
and change their measures. In this particular
example, to the current measurement they did.
They changed the blood pressure control level
so it matched the national measure of blood
pressure control. Is that something we should
push on for everyone? Is that important
enough to say that at least as they review the
evidence and we look at competing measures,
that even if we say it is fine, this measure,
the thing Patrick just rattled off that I
thought was great, it is clearly predictive of
outcomes. It clearly identifies patients
where they should go. It is valid. There is no bias. Do we still need to at least unpack to the extent of saying the evidence is sound and if there are related or competing measures, they are harmonized?

DR. DE LONG: Well you are asking two questions.

DR. BURSTIN: I know.

DR. DE LONG: In the harmonization, the reason not to is it is inconvenient because of historical precedent. But it is also inconvenient for people who are trying to deal with these and which one weighs more.

DR. BRATZLER: This is Dale. I think Helen, wherever possible it makes sense to try to harmonize the metrics, particularly where there is a strong evidence base to support perhaps the individual measure that you are comparing to. I just think it is really hard for clinicians to deal with competing measures.
MS. PAGET: Just one other comment on the unpacking and the science and the underlying pieces of the unpacking.

You know I think if we look ahead at where we are going from a technology standpoint and usability and accessibility of information for both patients, providers, institutions, and payers, we are going to have -- it is going to be easier and not harder. And so I think to move in a direction where we weren't offering the kind of transparency that people are experiencing in other aspects of their lives would probably be somewhat of a mistake. I think the transparency also about the purpose of the measurement and the pieces that went into the packing has to be really clear, particularly to consumers, so that we don't run into that problem with people wanting to identify one thing that may not have the scientific rigor that we wanted to.

But I do kind of come back to that Consumer Reports model and I think your point,
Helen, is right on. I mean I often find myself saying well that piece is, I'm okay that they didn't score that high there but I am really happy that they got this other score. I mean I have got a kid I hope is going to go to college next year. And you know you go through all these numbers but it is when you piece them apart that the tool that you use becomes much more advantageous, I think, in that kind of decision-making.

So I just think if we keep in mind where we are headed around how we are using our phones and everything else that that piece is going to become more important.

DR. ROMANO: If I am sensing where we are, so I think that what we are saying is that certainly NQF should continue to endorse measures that contain measures that aren't separately endorsed. And if they are separately endorsed, that may streamline the process a bit because you may be able to rely on the validity evidence in particular that
was presented through that separate endorsement process to say that these components are assumed to be valid as components, not for the higher order construct but as components. They are assumed to be valid.

But if there are other measures that are rolled up into a composite that are not NQF-endorsed, then what do measure developers need to do? So do measure developers need to present evidence that those other unendorsed components are in and of themselves valid or do they simply have to present evidence that those other unendorsed components contribute favorably to the overall composite?

DR. SHAHIAN: Going back to my work with the evidence task force, perhaps we could say that for component measures that were not NQF-endorsed that at least the developer would present the kind of evidence that normally would lead to endorsement such
as the quality, quantity, and consistency, and magnitude of net benefit of the proposed measure.

So parallel the kind of grading system that NQF uses, which is based on grade and USPS -- United States Public -- I always get that acronym -- you know what I mean.

(Laughter.)

DR. BURSTIN: USPSTF.

DR. SHAHIAN: But at least have some sort of explicitly defined criteria by which those non-NQF-endorsed components could be evaluated.

DR. BIRKMEYER: I agree with everything that has been said but I think it is also important to not lose sight of how heterogeneous these measures are, in terms of how they are put together and how they are used. And I think ultimately we might want to insist that we grade each of like several measures that are put forward, have them get graded in each of three or four different
domains. But at the end of the day, there needs to be like some holistic judgment that sort of rates those component grades against how it is to be used and it would be very analogous to how we rate grants that are submitted to the NIH, get a letter grade for the significance, a letter grade for the innovation and environment and methods, et cetera, et cetera. But at the end of the day what determines whether it is in or out or where it ranks is the impact, which is not just an averaging of those letters. It is basically weighting them according to a variety of other factors.

So a composite approach to scoring the composites.

DR. KAPLAN: Sometimes the evidence also comes from different levels. And now there is a new purpose. So now you have got these diabetes measures that we know some of the hemoglobin Alc, for example, predicts photocoagulation somewhere down the way. So
we know at the patient level that that is what these individual components do.

What we don't know is when you add them all up are you evaluating physician components? What components of the variation belong to the patient? There is nothing the doctor can do about that. What components of the variation belong to the doctor?

So are you asking, and I am just throwing this out, when you are using it for a different reason than the evidence substantiates its intent for, are you now going to challenge the stewards to provide evidence that actually it is okay? What is the reliability and validity for use at a different level than the evidence supports?

DR. ZASLAVSKY: Just to elaborate on what I think both John and Sherrie were saying. I think that often as a practical matter, measures are developed based on a limited amount of data in some pilot study that someone has done. So you go out and you
test something in seven hospitals or with 30 physicians or whatever. And you come up with a model that looks like it is pretty good, it seems scientifically plausible and so forth.

But now what you would really like to know is whether you can use this to make comparisons among hospitals. And you can't say anything based on the ten hospitals you are able to recruit to your pilot study about whether it really predicts outcomes at that level.

So in that case, I think you need to give a kind of a qualified or kind of provisional approval that says on the evidence we have, there is enough here that makes it worthwhile to do what is a substantial investment people are going to have to do before it is fully validated at the level that you want to validate it at.

And then take note of the fact that it is really a requirement that the data that further implementation that takes place
be brought back for either confirmation or improvements or maybe invalidation, you hope not, of what you did provisionally based on a more limited evidence base because you are not going to get those big evidence bases that you really need to be really confident that this is doing the right thing until you have actually given some level of endorsement to it.

DR. SHAHIAN: I guess I would just respond to that by saying that in 2012 and going forward, the stakes are so high that I would say it ought to be incumbent on the developer to do whatever needs to be done to prove that before they come to NQF. I mean, we just can't have more and more measures for which we don't have sufficient evidence.

DR. BURSTIN: So this is actually very helpful. It sounds to me that we are saying still yes, we should at least assess the evidence of the component measures. It is important to note that the Evidence Task Force
in our final guidance, at least, indicated that for outcome measures, there was only a requirement for a rationale. So it doesn't have quite as much and a lot of these wind up being outcomes. So I think that is just, I want to at least put that on the table.

The final bullet up here though is what I am hearing and tell me if I am hearing this correctly is that what is sort of at the highest level is the reliability and validity and the performance of the final composite performance measure but that whether or not we require additional analyses of the components is something that might be a decision that the developer may put forward to say this measure actually really only operates at the level of a composite performance measure. We don't necessarily think the individual ones do but they are important for the sake of the composite.

So I am getting the sense we are leaning towards saying no for the final bullet
there, additional analyses for the components, unless the developer believes those are stand-alone measures that should be evaluated for endorsement.

Am I capturing this discussion?

DR. DE LONG: I thought there was another issue which is are they necessary.

DR. BURSTIN: Right but necessary still to me seems like necessary to the final composite performance measure --

DR. DE LONG: Yes.

DR. BURSTIN: -- as opposed to individually.

DR. DE LONG: Right.

DR. BURSTIN: And we haven't done anything new on that.

DR. DE LONG: Okay.

DR. ROMANO: Just to make sure I understand this concept of necessary.

So are you talking about necessary conceptually or necessary empirically? There may be a distinction where you may have
things, for example, that don't contribute anything empirically because they are highly correlated with other things or because they are almost always done.

But people might argue from the face validity perspective that it is a fundamental component of what should be done. It is or from the outcomes perspective, that it is an important outcome. So even if it doesn't -- I'm just posing sort of a devil's advocate question.

DR. DE LONG: No, I think that is a good question. I am coming from the empirical evidence and the point of view that I believe we have kind of overshot the target in terms of how many performance measures we are getting people to record and report and look at. And to the extent that we can be more parsimonious I think that will be very helpful.

So if there are, for example, aspirin on discharge for an AMI, apparently
that contributes nothing. Is it important because it makes sense to include it in a composite when it doesn't add? I would say no but others may say yes.

DR. ROMANO: The question is does the developer have to show that it adds empirically or is it sufficient to show that it adds conceptually, in other words that clinicians, patients feel that this is important or do you also have to show that it adds empirically? That is the standard for review.

DR. KAPLAN: I was at a meeting of a bunch of physicians, a very large group of physicians and I was basically the lone psychometrician in the group but I was presenting the physician-level performance measures for diabetes. And it turns out that blood pressure outcome does not hang together with the other outcome measures empirically. So I said well we got rid of it. It actually improved the precision estimate.
We did a chrome block alpha. It improved the precision estimate at the physician level. I must have had 20 hands up exploding at me saying my patients are more likely to die of a stroke than anything if their hemoglobin A1c is in control for their whole life. So what are you doing throwing out blood pressure as a measure?

Well it was justifiable and actually very justifiable at the empirical level but there are times when it is just so entrenched in the credibility of a measure. If you leave it out, you do it at your peril. So just so you know, sometimes crass empiricism can really taking you down a dark hole and for NQF's sake, I would like at least the inclusion of the potential for if not empirically supported, at least the substantive rationale and leave it the way it is.

DR. ROMANO: Thank you for stating my concern more clearly.
DR. SHAHIAN: But the other side of that, of course, is when NQF endorses a measure, it goes to CMS. CMS rolls it out. It has a two or three percent impact on hospital reimbursement for all Medicare admissions. Hospitals push back and say I would like to see the evidence for that component of that measure. And you say well, people thought that that was important. That is the other side.

DR. ROMANO: Well perhaps some of it gets back to what the construct is. In other words, the clinicians who exploded at Sherrie's meeting were right in that stroke is an important outcome for diabetic patients and hypertension is an important risk factor for stroke and, therefore, controlling blood pressure has to be recognized as an important component of treatment for diabetic patients. Even if it doesn't hang together empirically and add to the composite, it clearly is related to patient outcomes. So if we are
trying to pitch a composite as being useful for decision making by patients, how can we ignore hypertension?

DR. DE LONG: That brings up what was this measure for?

DR. KAPLAN: It's actually -- we published it and it was for assessment of physician performance of diabetes care. And so we took the NCQA diabetes recognition program data and sampled patients at the physician level and all that stuff and got these performance estimates at the physician level. That was supposed to be and it was used in whatever the acronym is for the recognition program in diabetes and NCQA. So it was to be used at the physician level.

The question is not whether or not high blood pressure at the patient level is a bad thing to happen. The question is whether used as a performance measure for estimating physician performance, it contributes, and it does not.
So if you look at the physician effect on blood pressure, forgive me for the clinicians, it is buckshot. There is no signal for physician effect on the blood pressure outcome measure.

There is a signal for glycated hemoglobin. There is a signal for LDL. But there is no signal for the blood pressure outcome and that is why it wasn't able to contribute to the variability and physician performance measure.

So it kind of depends. I wouldn't put anybody through that analysis. That is way in the future. But for the sake of, and because I got myself into terrible trouble by excluding that, for credibility sake, and I appreciate the issue of credibility versus evidence, I think they are going to be a tension going forward in how we put these measures together until we are much more sophisticated in our audience.

DR. DE LONG: Not to dwell on a
minor point but the physicians were up in arms because you weren't measuring was it blood pressure control?

DR. BURSTIN: Yes.

DR. DE LONG: Even though they have no impact on it?

DR. BURSTIN: Well I think that is debatable.

DR. KAPLAN: Apparently they think they do and empirically it didn't show up.

DR. BURSTIN: Dual versus group.

DR. KAPLAN: Yes.

DR. BURSTIN: Of course. I do impact blood pressure, I am convinced of it. So I think that is a tough example.

I do have one question that raises for me, though, which is that we do oftentimes as we get into the criteria discussion this afternoon, we have rating scales for some of these things. So one question might be for the composite do you get extra points if you do both empiric and conceptual. And there may
be it is modern and acceptable if you do one
or the other but to David's point, if these
really are sort of some of the higher stakes
measures, people may not be happy having it
just be one or the other. Moderate may not be
good enough going forward. But that may just
be one way to do it. So quantity, quality,
consistency it is as clear as could be that is
high. You have to have consistency. That is
a no-brainer. You are low if you don't have
consistency. And maybe there is sort of a
similar construct we need to think through
about what is the requirements for the
composite.

DR. WRIGHT: But that implies
value. So you want to be clear about that
that you are evaluating one type over another.
And if that is what you want to do, fine.

DR. BURSTIN: Another thing
valuing one or the other, the question is is
there value of having both that is additive
beyond one or the other. I'm not saying -- I
wouldn't even know how to pick between the two
of them. But is there added value, if in fact
you have demonstrated both?

    DR. WRIGHT: Yes, both would be
fine but I would want to differentiate between
the two.

    DR. DE LONG: And once again, you
are measuring physician performance. There
are outcomes like stroke that are being sort
of also evaluated among these patients.
Wouldn't -- do we want blood pressure in both
of those measurements -- measures?

    DR. KAPLAN: This is where you get
into the weeds. I think it is probably not
worthwhile pursuing all this stuff. But the
point was if there is a clinical rationale for
including something versus an empirical
rationale, maternal mortality is probably a
better example. It never happens but when it
does, it is really bad.

    So even though probably
empirically it wouldn't make it into a measure
of hospital performance, clinically it is so
devastating, you had better put it in there as
a reflection of -- you know, that is just an
example. But I think sometimes we get into
the crass empiricism way too much.

DR. ROMANO: So lunch is here.

Should we -- okay.

DR. BURSTIN: Do you want to just
see if Dale wants to say something?

DR. ROMANO: Yes, so I think we
are reaching a breaking point in our
discussion.

(Laughter.)

DR. ROMANO: Dale, would you like
to add anything at this point?

DR. BRATZLER: No. It has been a
fascinating conversation.

DR.-ROMANO: Okay, should we open
to public comment, then? Is there anyone else
on the line? Can we open the line?

DR. BURSTIN: And also in the
room.
DR. BURSTIN: Monica, can you open the lines for us for public comment, please?

OPERATOR: At this time, I would like to remind everyone in order to ask a question, press *, then number 1 on your telephone keypad. We will pause for just a moment to compile the Q&A roster.

At this time, there are no questions.

DR. BURSTIN: Thank you.

MS. CRAWFORD: Thank you. I just wanted to speak really briefly.

Thank you. My name is Alyssa Crawford. I am here from Mathematica Policy Research and we have a number of measure development projects with a number of other partners and for a number of agencies.

I just wanted to speak up really briefly because I think a lot of the discussion in the past 30 minutes, in particular, has applied to a lot of the things we have discussed internally. And I wanted to
1 put out a plug that I think there is a
2 difference between a lack of evidence and
3 evidence of a lack of effect when it comes to
4 scientific acceptability. And to some extent
5 when it comes to evidence, sometimes there is
6 just a lack of evidence to support whether a
7 concept is actually improving outcomes and
8 that doesn't necessarily mean it doesn't
9 improve.

10 We have seen this in particular in
11 certain measure development areas such as
12 behavioral health, where there is just not as
13 much evidence out there to support. And there
14 aren't as many measure out there in the field
15 for people to choose from. So I just wanted
16 to bring that to the group for consideration.

17 DR. BURSTIN: I'm sorry. Lack of evidence versus -- could you just do the
18 second --

19 MS. CRAWFORD: Sorry. I think there is a difference between evidence of a
20 lack of effect and a lack of evidence of an
effect. So whether or not you have evidence
to prove that there isn't reliability and
validity, versus the evidence doesn't
necessarily support it.

DR. BURSTIN: Right. And NQF has
an evidence exception specifically for those
areas that we invoke as necessary. We try not
to use it very often but specifically to say
we recognize there are times when the evidence
is just not there yet. We don't invoke it
very often, for example, spiritual care and
palliative care. Again, weighting for those
studies seem unnecessary and we specifically
allow the committee to put forward their
expert judgment and invoke the exception to
say in this instance, the benefits
significantly outweigh the risks to patients.
But thank you.

DR. ROMANO: Okay, so not hearing
any other questions, we will break for lunch.

What time will we come back, for
Dale's benefit? One o'clock. So we will
reconvene at one o'clock.

DR. BRATZLER: Okay, I am going to hang up and I will call back in.

(Whereupon, the above-entitled matter went off the record at 12:20 p.m. and resumed at 1:10 p.m.)
DR. ROMANO: Okay, so we will reconvene. Thank you again for joining us, those of you who are on the phone. I think our task for this afternoon really shifts into the weeds, where we go from kind of a broader conceptual discussion into looking at the specific measure evaluation criteria. Do we have a side on that by the way? We probably shouldn't keep that slide up.

So I think all of you in your packets have a number of relevant materials. So you have the measure evaluation criteria from January 2011 that are currently used by NQF and you have a table that summarizes, it is called Table 1. It summarizes individual and composite measure evaluation criteria. And this is a nice table because in the left panel it highlights the criteria that are used for evaluating individual measures and then on the right panel, it shows how those criteria
are altered or added to if a measure developer declares that they are submitting a composite measure.

And then there is another sort of 12-page document here that represents what measure developers actually have to fill out currently for composite measures. We will not go through the 12-page version of this although some of us have done so but it is here for your reference and the message is that each time we say that there should be a certain criterion or a sub-criterion, that NQF staff has to convert that into a box, an item on this form or a set of boxes or items that measure developers then fill out.

So if you have thoughts about how that should be done, please feel free. We won't be going through the form in detail but please feel free to offer your thoughts as we go through the discussion.

Now I think Karen and Karen was it also came up with this table that is part of
the first agenda packet, the briefing memo. It is called DRAFT Table 4. So I would ask that people sort of have in front of them -- okay, we have it on the screen as well. So there is DRAFT Table 4. So this shows the current additional criteria for composites and considerations that have been raised by NQF staff and through previous NQF steering committee processes.

So what we will be doing is we will be going through Table 1 that is in your packet with the current individual and composite measure evaluation criteria. We will be going through the considerations that are shown here for this DRAFT Table 4 and we will be making recommendations to NQF about specific wording changes, specific operationalization.

So to start in Table 1, just to give everybody a frame because some of you have gone through the process as developers or as Steering Committee members but others may
not have. So there are four conditions for consideration that must be met before proposed measures may be considered and evaluated for suitability. This is true across the entire enterprise.

The measure has to be in a public domain or an intellectual property agreement as signed. There has to be a responsible entity and process to maintain and update the measure. The measure has to be intended for both public reporting and quality improvement and those terms are defined broadly. So it could be public reporting in any context, quality improvement in any context. And then D, the measure submission itself has to be complete.

So those four considerations I don't think would be altered here because they apply across the enterprise. Right?

DR. BURSTIN: Right.

DR. ROMANO: Okay. So we will go into the next -- everybody with me?
So we will go into now the first of the four formal criteria for measure evaluation, that is, importance to measure and report. So importance to measure -- oh.

Okay so before we do that. So I think that we have agreed this is on what is labeled as page nine of this Table 1, criteria for evaluation. So currently composites have this higher level criteria for evaluation that says that the individual measures included in the composite must be either NQF-endorsed or assessed to have met the individual measure evaluation criteria as the first step in evaluating composite measures. So I think we have agreed that we are actually dropping the second part of that or statement.

Can we put that up on the screen? Do you have that? Okay, let's put it up on the screen just so everybody is clear about it.

Because this is kind of the first step. So what Liz and I heard, I think, and
Karen and Helen from this morning's discussion was that rather than requiring every component of a composite to be either NQF-endorsed or to have met the measure evaluation criteria for endorsement, what we are asking for instead if the components are not NQF-endorsed is that there be some evidence for its inclusion in the composite. If we could scroll down a little bit. Okay.

So the second part of this or statement would be that there should be some evidence for the inclusion of components that are not already NQF-endorsed. And that evidence could be either based on the -- or probably both. Maybe it is both. But it should be based on the individual performance characteristics of the component, particularly validity as Alan pointed out earlier. Reliability may be completely immaterial for a component of a composite but there may be evidence based on individual characteristics.

Next page. There we are. Thank you.
Okay, so rather than saying or assessed to have met the individual measure evaluation criteria I think what we are looking at is or there is evidence for its inclusion in the composite based on either the individual -- its performance characteristics and individual measure or based on its contribution to the performance of the overall composite.

That contribution could be described, in most cases empirically but in some cases it might be described conceptually as we discussed with Sherrie's example of blood pressure this morning.

Does that capture the sense of the discussion in the latter part of the morning? Okay, so we are fairly fundamentally changing the second part of that or statement.

Okay. So given that, then let's move on to criterion one, which is importance to measure and report.

DR. SHAHIAN: So just before we go
on, just in terms of the optics of what we are doing, we just went through a process two some-odd years ago where we tightened the evidence criteria and made them much more explicit. Will this be viewed as a weakening of the evidence criterion? I'm just asking.

DR. BURSTIN: It is actually interesting because I thought we were going to just go to the criteria first and then try to figure out where we are on the bigger one because it is a little fuzzy for me still. Because I think what we said earlier was we still wanted evidence until we get to the next criteria. We are still expecting evidence and, I would argue, performance gap as well for the components.

So we may need to nuance this wording a bit.

DR. BIRKMEYER: But I think that while I agree that we shouldn't roll back the tide with regards to evidence, I think that the focus should be on sort of the evidence
around how well the summary score works in the context of what we are talking about here, rather than each component that rolls up into it.

DR. ROMANO: So in a way we may be lowering the bar for components but we may be raising the bar for the composite as a whole and for ensuring that the construction of the composite is based on a clear construct, a clear quality construct.

DR. BURSTIN: Does that work for you?

DR. BIRKMEYER: I think so.

DR. BURSTIN: Okay. Since you chaired the evidence Task Force.

DR. ROMANO: Okay, so let's look at the importance issue. And so what was done before, three years ago was it, was that criteria 1a, b, and c were retained but new criteria 1d and e were added for composites.

So 1a is about a high impact aspect of healthcare; 1b is about a
demonstration of opportunities for improvement; and 1c is about the evidence base, if you will. Is that an appropriate summary? Okay.

So the additional evaluation criteria that were added for composites, 1d was that the purpose/objective of the composite measure and the construct are clearly described and 1e is that the components are consistent with and representative of the construct.

So could everybody just read 1d and 1e as we are talking and think about how those should be changed or adapted, based on our discussion?

Basically what NQF staff said here is that this has been difficult to apply in practice. These criteria 1d and particularly 1e. 1d would seem to be relevant to every performance measure and not unique to composites but I think we have already said that there is a blurry line and probably most
of the measures that are endorsed by NQF are composites if you really look at them closely. And 1e seems difficult to apply in practice. So any thoughts or responses to those considerations?

DR. ZASLAVSKY: It seems that the piece of 1d that might need to be made more explicit is that the method of forming the composite has to be justified by reference to the objective of the composite measure and the conceptual basis of the composite measure. That is what we were talking about this morning and it isn't really in that -- that is what is different for a composite as opposed to anything else.

DR. ROMANO: The method of forming the composite from the components has to be clearly linked to the purpose and objective of the composite measure. Is that what you are saying?

DR. KAPLAN: Yes, I was missing the appropriate part. And they have to be --
the purpose has to be described and it has to be appropriate.

DR. ZASLAVSKY: Not related by virtue of being involved with this.

(Laughter.)

DR. ROMANO: Anything can be described. So you can describe a method for adding together apples and elephants.

DR. CHASE: So let me test this forward because I can see where this gets difficult. Again, if a purpose could be well there is a whole bunch of criteria that we have collected and they are individually valuable, it makes them easier to understand if we combine. Then just about everything people are going to bring is going to meet this. Or are we saying that is never a reason to do composites and yet I think that is a reason why people are doing composites sometimes.

Again, take a prevention composite, which would stand alone on each one
of its components and I might well want to
bring it just because why give people nine
things to look at when they could look at just
one.

DR. ROMANO: Well let me just push
back a little bit and say that making it
easier is not clear enough. That you have to
say making it easier to do what. Making it
easier for making what decision that is
relevant in this marketplace?

DR. KAPLAN: Yes, I mean the
purpose -- I would separate those two. The
purpose has to be explicitly articulated. The
rationale for creating a composite has to be
explicitly articulated and then the
appropriate of use for whatever construct for
the purpose of is appropriate.

So I would separate the two
things. One, you have to have -- the purpose
has to be explicitly articulated and then the
methods and the construct have to be
appropriate for that purpose.
DR. ROMANO: So then what I am hearing is that 1d has essentially two subcomponents, where the first subcomponent is related to explicit articulation of the purpose and the second subcomponent is related to how the methods follow from that purpose, the appropriateness of the methods based on that purpose. Is that what people are saying? I see some nodding.

DR. BURSTIN: It might be helpful to actually just to use an example. Let's keep on Jim's example for a moment.

So somebody takes all the currently endorsed NQF-endorsed measures around prevention and screening. Let's just make it easier, just a screening composite and brings it forward. What would need to be -- just give me a sense of what you think would be an acceptable explicitly articulated purpose for a screening composite.

DR. CHASE: My argument would be again when I said ease, I meant for a consumer
to look at one indicator of the overall
prevention that a given provider organization
provides.

DR. BURSTIN: That works for me.

I'm just curious if it works for everybody else.

DR. CHASE: And later in the process might come the test of is that a valid thing to do, to combine all seven for performance.

DR. ROMANO: But I think what is maybe needs to be a little bit clearer again relative to what developers are used to doing is that this is intended as a measure that consumers could use to choose physician organizations, provider organizations that provide a higher quality care in prevention and screening.

Is that -- yes. So that makes it different, for example, from a measure that -- so you could formulate a different composite where some of the words would change but it
would have a very different construction
because it would be designed for payment
determination, for example, for providing a
financial reward to provider organizations
that are improving patient outcomes.

DR. BURSTIN: -- for specific
purposes so that gets funky.

DR. ROMANO: All right, I
understand that.

DR. KAPLAN: Well like --

DR. ROMANO: You can't avoid the
problem.

DR. KAPLAN: American Board of
Internal Medicine did these performance
improvement modules and they created
prevention, chronic care and acute care as the
performance things they were trying to
evaluate. Well the prevention thing didn't
work so well. It doesn't hang together too
well. The acute care not so well but the
chronic disease care worked really well at the
physician level.
So the point of those measures was to create a composite out of the things they have already -- you know that are already around that looked at the doctor's ability to provide high quality chronic disease care for their patients. So that was the underlying purpose and then they combined all these measures in ways that we helped them with to provide the empirical, the evidentiary support that actually those measures were appropriate for evaluating physician performance in terms of the chronic disease care provider.

So that is another -- it is a different example but it is another kind of way of looking at if you have got a new purpose out there, you are going to combine these things differently and then we are going to ask you to state what that purpose is and how you are going to do it.

DR. ROMANO: If they don't -- let's take that example. If they don't hang together empirically, then what do you say?
Do you go back to Jim and say you cannot do this because they don't hang together empirically or do you say you can do this but consumers ought to know that prevention for men's health doesn't necessarily correlate with prevention for women's health or prevention related to breast cancer screening doesn't necessarily correlate with prevention in other domain. How do we respond?

DR. KAPLAN: Well I am not NQF. So how I would respond is differently. I would say those measures that you just handed over don't look to be good measures of physician performance. They may be very good measures of planned performance or they may be very good measures of patient something or other but they don't look like they are very good measures of physician performance unless you have more of the same things. In other words, as a measure of physician performance, this may not shake up.

Now, should that paralyze us from
never doing physician performance levels of prevention or are we at such a crude state of understanding this process that we will take whatever it is they offer us up because we are not sure that they are not very good measures of physician performance, for exactly the reason I stated. There may not be enough of them. They may have too much patient variability. The doctors may be attracting patients with certain kind of wellness profiles and so on.

But I would at least like to hear some language in there about here is what you are trying to do. Here is what your intention is, and here is the methods you are setting about to accomplish that.

DR. BURSTIN: It feels like we are blending two criteria. So for me at least there is the evidence, which I think is different. This one about impact evidence and opportunity for improvement feels different to me than the empiric basis that we are now
talking about.

So I guess one question is going back to David's comment that if we are raising the bar on the composite overall around evidence, I mean is this essentially the conceptual piece you were talking about earlier that you have to really be able to provide the evidence for the composite conceptually and then the more empiric assessment winds up in the next criteria outside acceptability where you actually show the data on reliability and validity for the composite?

DR. DE LONG: Can I have some clarification about evidence? When we talk about evidence, are we now talking about evidence for the composite for what it is doing and not for the components? Because that is a whole different set of evidence.

DR. BURSTIN: Yes, we are talking about the composites.

DR. DE LONG: All right.
DR. ZASLAVSKY: I've been grappling with this question that Patrick posed of what the bar is for a composite which didn't have an immediate answer. And I think I was trying to figure out what the bar is for something which isn't going through the composite process. And it doesn't in that section anyway say anything about showing that it is useful for a particular purpose. It just says it has to be a measure. Maybe I am missing something there.

DR. BURSTIN: You are absolutely right. Useful is really about usability. That is a different criteria.

DR. ZASLAVSKY: Yes, and usability doesn't actually necessarily imply usefulness either.

DR. BURSTIN: Actually now usability and use which is intended to imply usefulness as it has been changed.

DR. ZASLAVSKY: But anyway one place I came to is maybe we do have a higher
standard for a composite to get an NQF seal on it. Anyone can take a bunch of numbers that come out of measurement processes that have been validated and so forth and throw them together any way they want to and put it on their reports and maybe even get paid for it. But I don't think that the fact that they bring that in front of NQF and the components are all okay necessarily means that NQF wants to say anything, give any kind of approval to that.

I think that there this additional step of creating the composite that NQF is being asked to approve and that it is reasonable for us to ask that there actually be value added in that step as evidenced by there being thought about what conceptually it is getting at, what its purpose is and the appropriate kind of evidence to meet the standard required for that purpose.

So I feel okay about having these additional requirements which go a little
beyond what we do with the individual measures.

DR. DUNTON: Can we step back to the purpose discussion for a minute? If we have to have measure that are for all purposes, and in usability we have to describe how that works, can we narrow it down to one into such a statement for a specific composite?

DR. BURSTIN: You know again, this is really in a different section. This is about is this important to measure and report. So I think it is fair game for the developer to put forward their conceptualization that this measure be especially important for the following uses.

But again, that doesn't mean that measure will only be endorsed for the specific uses, I guess.

I would be curious David as you think about, just because you have been through this, is you think about the
individual CABG measures you have already
developed and then you think about the CABG
composite. If you were filling this out, what
would you say that would be higher? I can
tell you are smiling already. I mean to me,
that is the question. Is this something
substantial or is this something that sounds
like a little mom and apple pie?

DR. SHAHIAN: Well this is a
little scary, I think you reading my mind
because I was thinking about the CABG
composite and I think it is probably going to
be true of many composites.

There is an incredible amount of
evidence out there for the individual
components of the CABG composite and there are
zero evidence per se that a composite of CABG
measures makes any difference.

So the evidence was solely at the
level of the individual domains and measures
but not at the level of the composite. And I
suspect that that is going to be true in many
cases because we don't have a lot of testing
on composites. Is that what you were getting
at?

DR. BURSTIN: Well is it really
that it is maybe evidence is again we are
trying to pound this square peg into a round
hole. Are we really talking about evidence
for the composite or are we saying maybe it is
really impact? So what is the added impact of
having these measures in a composite versus
individual? Maybe it is evidence -- we are
really talking about evidence for the measure
focus. Evidence, I am not quite sure it is
the right --

DR. SHAHIAN: And in the case of
the CABG composite the reason we devised it --
well there were many reasons. One is that it
is increasingly difficult to distinguish
levels of performance based on mortality
alone. So that is number one.

Number two, there was when we
developed this five or six years ago an
increasing realization that quality measurement should be multi-dimensional. And we had one very narrow dimension of quality that we were measuring. And this was a way to incorporate mortality, morbidity, process measures.

And then number three was consumer interpretability.

So there were many reasons that we did it and absolutely no evidence prior to the introduction of the composite that it was really a good thing.

DR. BURSTIN: And to me those three are quite strong conceptual reasons why we would have a composite. Maybe it is really a conceptualization rather than evidence.

DR. ZASLAVSKY: The nature of evidence is pretty broad because we have a lot of alternative conceptualizations. So the conceptualization here is maybe that these are all things which are either measures of outcome that affect patient well-being for
which there is evidence that the processes are things that contribute to those outcomes. So that is an offset of arguments. It is not a set of the empirical exceptions so as far as you refer to the empirical evidence that the processes have contributed to outcomes. But it is information that some of them put together and say this is a reason for this to be a composite, rather than just saying I group together ten things arbitrarily and say put a label on them.

DR. KAPLAN: To me this is still winding around the issue of purpose. Because if collectively these things are telling you something new than they would tell you individually, it is a different purpose you are putting them to.

So now I want to evaluate physician performance or I want to evaluate the hospital performance on a construct I could measure individually but collectively, these things tell me something more robust.
And that new whatever, you can call it a
higher order construct, you can call it a
robust composite. You can call it something
but it is something else that hasn't been,
that these individual components don't tell
you. This new collective enterprise, whatever
construct, whatever you are calling it, the
new higher order thing is telling you
something different and it is a new purpose.
Now you are putting these things to a new
purpose and the purpose is blah, blah, blah,
blah, blah and here is why that is important,
if you are leaving it to the impact.

That is important because we don't
have right now a good measure of this new
blank to estimate whatever it is we are trying
to estimate.

DR. CHASE: So --

DR. BURSTIN: I agree with whatever
you say except the word purpose is throwing
me. That's all.

DR. CHASE: So can I -- I just
want to devil's advocate on this. We talked a lot about how this would be used but is this one ever going to screen anything out?

DR. BURSTIN: That's what I was asking.

DR. CHASE: Can somebody give me an example of where you wouldn't be able to just for just about anything? Because again, at some level it is just I am putting it together so your point earlier, so consumers will find it easier to see and to use.

And when you take many things and put them into one, unless you go through the other pieces where they are invalid, those individual pieces or invalid or one of them doesn't add to it, I am just --

DR. KAPLAN: I guess I would have said what are you measuring. If I am measuring physician performance, is this a new measure?

I mean you can add up everything. Don't limit it to the whatever 50 measures of
safety you have or whatever. Add up
everything NQF ever endorsed ever and what
have you got? You know, that is ridiculous.
You can't start at it that way. You have to
have some construct you are trying to
estimate.

So what are you trying to measure?
And then does this new composite thing reflect
that better than individual elements do?

DR. ROMANO: Yes, I might argue
that if people are really forced to describe
their thinking here clearly, that it will open
them up. It will at least foster some robust
discussion in steering committees that may
lead -- for example so if we are talking about
diabetes care, not to pick on you
specifically. But so what is the concept
behind -- is optimal diabetes care? Is that
the measure? So what is the concept behind
that measure? So you could say well this is
to facilitate consumer decision-making. But
then I might say well if you want to
facilitate consumer decision-making, getting back to Liz's argument, shouldn't that be done in a way that encourages consumers to lower their risk to improve their outcomes? And therefore isn't the implication of that, as we get later into the evidence, isn't the implication of that that a variety of measures should be included in a way that is all correlated with diabetes-related outcomes.

So it is not so much that people would reject the rationale is that people would then say well the construction of the measure may not be fully consistent with that rationale.

DR. CHASE: So the diabetes one I think fits here well in the sense of part of the argument with that one is those individual components getting all of the components are not just additive. It is they interact with each other so if your blood pressure and your LDL is in control, there is some evidence that you are going to be better off than just a
patient with LDL control only or just blood pressure control. We don't have to argue the veracity of that now. But I mean I can see where there are many cases where you would say absolutely makes sense to have a composite. I was trying to find the case of where you would say it doesn't. There shouldn't be composites unless you can actually show some argument about again why it makes sense to put them together and that is all we are asking.

DR. BURSTIN: Or maybe taking the example where we have actually looked at measures before that didn't make it through where you put together a whole lot of topped out measures and you still have got a topped out measure. Is that really adding value?

Maybe it does go back to whoever said it, but maybe part of this is actually weaving in the what are you trying to measure better than the individual elements on their own do? That goes to some of what I think David listed off. Maybe that is part of the
way to structure it because it has got to be better.

Let's just try the higher bar.

Why is it better than simply taking the individual measures? And obviously I think the real tough part is going to be when we get to the testing. So you probably I think --

DR. ZASLAVSKY: Patrick had a good reformulation. The rationale isn't to make it easier for consumers to look at it, it is to make it easier for consumers to be informed by it. So the composite has to be informative. It has to be able to convey valid information, which is not just a hodgepodge of stuff thrown together. It has got to have some thought behind it that this is a good thing for them to see.

DR. ROMANO: And the implication then is that it sets up the next stage, which is evaluating the evidence because the evidence then is evaluated in the context of the developer's stated purpose and in fact is.
DR. BIRKMEYER: That's right. I was just going to say the same thing that I view 1d as not a screening tool to identify applications, this shouldn't be here, but a way to reinforce what the primary purpose is to make it easier for the reviewers to judge the ultimate value.

DR. ROMANO: Okay, so given that, so there will be, obviously rewording that I guess NQF staff will work on and we may discuss that in a subsequent conference call.

So anything else that we should discuss in the context of evaluation criteria and one here for importance? One thing maybe that struck me in looking through this is that 1c becomes a bit difficult to answer in the context of a composite. 1a is about is it high impact. That is usually sort of a qualitative argument that is fairly easy for people to make, although sometimes people fail in it.

1b is about variation. That is
usually empirically demonstrated with in this case the composite measure as a whole showing the composite has performance variation, opportunities for improvement.

But 1c is a bit awkward because some of our composites actually combine process and outcome measures. So how would people answer 1c for those types of composites?

DR. BURSTIN: I think we said earlier they would have to look at the individual components evidence.

DR. ROMANO: Their 1c wouldn't apply then.

DR. DUNTON: Well just add a category for both.

DR. ROMANO: Well should -- let me -- I think it was Karen posed this question. Should there be for all measures, not just self-declared composite measures, should there be a 1d statement that developers are required to make about the purpose or objective of the
measure? Forget composite. Not really, there is a description.

DR. DUNTON: Okay, it's a good point.

DR. ROMANO: What do people think? Should this be something that is required of every measure?

DR. DUNTON: It is there in 1b, really.

DR. BURSTIN: I was going to say that.

DR. ROMANO: It is implied in 1b.

DR. BURSTIN: It would be fine to have that explicit and then the additional burden for composite is and how is that better than the individual ones. That is fine. We can consider that.

But could we talk about are we still on -- and we talked about this earlier that we did think that for the components of the composite that they needed to be evidence-based. So the evidence for the measure focus
should be there, particularly also to allow us
to then compare existing measures in the
harmonization issues.

DR. BIRKMEYER: But again there
the evidence means that there is either
evidence of good performance as a stand-alone
measure or evidence that it contributes to the
performance of the composite as a whole.

DR. BURSTIN: No, this isn't
evidence for the measure focus. This is
literally the quality, quantity and
consistency of the evidence for the measure
focus. Is there evidence that blood pressure
of 140/80 is the right number, if it is one of
the components?

DR. KAPLAN: When you get new
measures, something that is actually going to
make the composite more robust and there is
good expert opinion that that is what is going
to do. But there isn't good empirical support
for it but the measures developers can create
a rationale such that you are adding things
that actually reflect whatever it is your
performance at the physician level,
performance at the hospital level, whatever it
is you are trying to measure. These things
are conceptually very good contributors to
this, we think they are. Are you going to
stifle, are you going to cause problems here
in creating composites that are better and
more robust by limiting it to an evidentiary
base for some different purpose or some
different level?

DR. BIRKMEYER: I thought we
talked about this earlier at length.

DR. BURSTIN: I did, too. I
thought we had actually when we had this
discussion -- maybe I am off base but I
thought when I talked about this earlier, we
did say that the components within the
composite should pass the evidence test, I
thought. That we weren't going to require
them to have individual testing and they could
definitely be not reliable on their own but,
personally, as a clinician I wouldn't feel very comfortable that there is measures within a composite that are not evidence-based. I mean we are going to get huge push back on that. Evidence, not that it adds to the composite but that evidence backed, particularly if it is a clinical issue is evidence-based, unless it is an outcome, for which case you just need a rationale for why it is appropriate. It is more so on the process side that I think it is an issue.

DR. BIRKMEYER: On the process side I can see that this is the discussion that we had earlier with regards to the example that Sherrie had around a clinically credible measure that if you take it out, it just deflates sort of the oomph of the broader measure.

DR. ROMANO: Okay so if I think what I am hearing is that with respect to 1c and this idea of the evidence base according to the type of measure, that if it is
composite that includes process measures, then each of the processes within that composite should meet the evidence criteria. Whatever it is.

So each of the components within the composite should meet the criteria that are relevant for that component.

DR. BURSTIN: Yes.

DR. ROMANO: Sherrie says no.

DR. KAPLAN: Well not necessarily for the new purpose you are putting it to. Because there may not be any evidence that that contributes to the new purpose, just like the blood pressure example is important at the patient level. We know that. But for estimating physician performance, it doesn't contribute.

So it depends --

DR. ROMANO: You are talking about evidence for a purpose.

DR. KAPLAN: For a purpose, right.

DR. ROMANO: She's talking about
DR. KAPLAN: That's what I'm trying to clarify.

DR. ROMANO: -- rationale sort of.

DR. BURSTIN: And that is what 1c is. It is evidence for the measure focus.

DR. KAPLAN: All right.

DR. ROMANO: Okay, so are we okay, then? If the evidence for the measure focus is it actually better to have a blood pressure of 140/90 than to have a higher blood pressure?

DR. BURSTIN: Right.

DR. ROMANO: That is an intermediate outcome measure.

DR. BURSTIN: Yes.

DR. ROMANO: Currently, NQF would require evidence from clinical studies that that is a good thing and that should still be required if it is included in a composite. Right?

DR. BURSTIN: Yes.
DR. ROMANO: Okay.

DR. BURSTIN: What about the gap or variation? Gap in care or variation, would you require that the individual components have a gap in care or variation? You would be fine with topped out measures in a composite if justified?

DR. CHASE: Well we gave the examples where maybe for -- you might include it where it is important for patient communication. We brought that up as an example where there is not a lot of variation but you want patients to still know that that is an important thing to do.

DR. BRATZLER: This is Dale. So I think you have to be a little bit cautious about topped out measures that don't discriminate. Particularly, it gets to some of the experience we have had with the all-or-none measures. But if you have topped out measures particularly that have a big denominator, they can make your composite not
particularly valid. So you just have to be cautious about including topped out measures.

DR. KAPLAN: This is Sherrie. There is an old saying in measurement science: you don't measure what doesn't vary. So you wouldn't want to measure -- in the old days you wouldn't want to measure diversity using gender in the VA. But now the reverse is not true, however. For floor effect problems, things like maternal mortality. If it happens, it is so bad that you need to include it, even though the variability is so limited that you are not going to be able to use it alone. And this is what I think we were talking about before. It is insufficient by itself to constitute a quality indicator but collectively it could contribute to an overall quality indicator because when it happens, it is so terrible. So we wouldn't want to put the same criterion on a floor effect problem.

DR. ZASLAVSKY: So I think that means a modification of 1b for the measures
going into a composite.

DR. BURSTIN: And I guess it depends on what is the modification. I am hearing, tell me if I am hearing correctly, that in general you agree there should be a gap but there may be extenuating circumstance that perhaps you could justify inclusion of a measure as part of a composite. But I think it would require some justification. Does that sound fair? Okay.

DR. ROMANO: Yes, I mean it is also important to keep in mind, I mean Dale raised an important point but that can be dealt with through appropriate weighting. In other words, if there isn't undue weight put on the topped out components of a composite, then they can be retained without skewing the overall results of the composite, if it is important to do so for conceptual reasons.

DR. ZASLAVSKY: I actually meant to refer to 1a, high impact.

DR. ROMANO: Yes, I think that --
DR. ZASLAWSKY: You might want to soften that up. I think we have a real problem that you combine with a lot of other things into a composite in an appropriate way.

DR. ROMANO: I think we actually agreed 1a does not apply to individual components. Right? 1a only applies to the composite as a whole.

DR. BURSTIN: Yes.

DR. ROMANO: So 1a applies clearly to the composite as a whole. 1c still applies to individual components and 1b, as I am hearing were somewhere in-between. There are circumstances.

DR. BURSTIN: Got you.

DR. DE LONG: So I do think it is going to become relatively cumbersome to keep track of all of the topped out measures. I mean we are expanding at a fast rate here. And even if they make incredible sense, they are going to be taken for granted. I mean, they are automatically performed so they are
not performance measures.

DR. ROMANO: So I mean that is where people argue for all-or-none scoring, for example, that is a checklist and people should do everything on the checklist and if they miss anything on the checklist, it indicates a bad system of care.

We could argue about that but that is a rationale that is out there.

So let's move on to criterion 2, scientific acceptability. And scientific acceptability currently has a number of components. 2a has to do with the definition or specification of the measure, that is just very clear how it is defined and specified so that it can be implemented. 2a has been adapted for composite measures, basically to include components of how the composite is constructed. This seems reasonably straightforward. Any arguments about 2a, what should be added or subtracted from 2a?

DR. SHAHIAN: I think it is
actually quite well written. I like it.

   DR. ROMANO: Okay. So moving on then to 2b. 2b is about reliability testing. It is framed in terms of the repeatability of the measure results when assessed in the same population, in the same time period. And of course it references Footnote 8 which is about examples of inter-rater or intra-rater reliability, internal consistency, reliability for multi-item scales, test reliability for survey items.

   So how does reliability testing differ for composites?

   DR. SHAHIAN: We have the whole additional issue of inter-item reliability, which we don't really talk about here.

   DR. ROMANO: Well I mean I guess I would argue based on our discussion this morning that it may not matter. In other words if you are coming in with what you call a single measure from a CAHPS survey about physician patient communication, then you are
going to have to show the internal consistency reliability of that domain measure based on the construction of the survey. And similarly, if you are framing it as a composite, you are going to be showing the same thing.

DR. DE LONG: So what does it mean to show? I mean it seems to me that if you are going to use something to rate performance then you should demonstrate in some manner that you get the same ratings if you use it on, for example half of your data versus the other half. There has to be some consistency in the way this measure performs.

DR. ZASLAVSKY: Generally if you have done that for each of the components of the composite, then you can deduce that for -- especially since usually the different components of the composites often will be independent sources. If they are not, if you are taking two things off of the same survey, then you have to do the analysis where you put
them together. But in any case, the mechanisms for calculating that reliability estimate is going to be similar for the composite. So that is from the original --

DR. DE LONG: I would think that it would depend on the weighting of the individual components.

DR. ZASLAVSKY: Well it might depend on the weighting of the components. It might depend on the relationship about the measures. So it does require an analysis but there is nothing terribly different from that analysis from what you would do with a single measure that was a combination of different diamonds.

DR. ROMANO: Well so I guess where we get into some trouble here possibly is that this reliability concept is operationalized currently in different ways.

So for outcome measures, the way it is often operationalized is that the measure score -- is about the precision of the
measure score, basically the imprint, if you will, of the provider physician, the hospital in the case of the AHRQ measures. So we basically justify reliability based on demonstrating the hospital imprint, or in your studies the physician imprint, not based on internal consistency reliability.

So is everybody still okay with that? In other words that depending on what people have said in Section 1 about the conceptual framework for the composite, that may lead in different directions in terms of the reliability measures that are presented.

DR. KAPLAN: Yes, I mean it depends on the purpose you are trying to put it to and even the levels of reliability will tolerate -- I mean reliability is -- the question I was struggling with should NQF require standard reporting out of a kind of standard error of measurement or something that says here is the precision of this estimate for this purpose. Because for
example, in the certification process, ABIM really need -- you have to have a fairly, a really high bar of reliability because they are going to flunk somebody. So for that purpose, you really want to make sure that the estimate is very high. For big group comparisons, like I am going to compare specialists to generalists or somebody to somebody, large groups of folks, precision of the estimate may not have to be that big.

So if we are going to float this business about composites and purpose, maybe we should talk a little bit about what the tolerance is around error. You know, what are the consequences of making a mistake and what kinds of error can we tolerate?

Composites usually, you know falling on what Alan was talking about, composites usually buy you better precision. But usually with a composite you get improvements in precision, not reductions in precision.
DR. CHASE: Except I am worried about that sometimes you get improvement because you make assumptions about that the denominator is the same. I mean one of the things I worried about when we construct composites you take a prevention on an entire population and we do it from how many of the patients got everything they were supposed to but in reality the test is really -- I mean men don't get cervical cancer screening. So right there the real denominator is smaller. And then we just to make it easy we just assumed that the denominator is the full thing. And I don't know that it makes a difference probably practically but I do worry that we should be paying a little bit of attention when we do composites that we are looking at the reliability related to how the composite is put together, whether the denominators make sense.

DR. ROMANO: If there are no other comments on reliability, we can move to
validity.

2c is about validity testing to demonstrate that the quality of care provided distinguishes good and poor quality. And there are various levels or approaches to validity that are allowed in the measure evaluation criteria.

2b.1, 2b.2, 2b.3 -- so 2b.1 is about capturing the target population. 2b.2 is about the accuracy of the score and inflecting quality and 2b.3 is about the exclusions.

So any comments -- oh, and what is 2b.4? Oh wait, disparities is in here, too.

DR. SHAHIAN: Are you in Table 4 now in the other document?

DR. ROMANO: I am looking at the measure evaluation criteria table along with this. So 2b or validity is broken down into sub-components of validity.

DR. BURSTIN: Right, page 11 of the other document.
DR. ROMANO: Page 11 of the other document.

DR. SHAHIAN: It is actually b sub 1, 2, 3, 4, 5. Right?

DR. ROMANO: Correct. Right.

DR. SHAHIAN: Okay.

DR. ROMANO: Yes. Okay, so if you look at the left-hand column on page 11 there, 2b.1 is specifications consistent with evidence; 2b.2, validity testing for data elements or the performance measure score; 2b.3, justification of exclusions; 2b.4, justification of risk adjustment; and 2b.5 identifying differences in performance; and 2b.6 comparability of multiple data sources.

So issues in how these differ for composite measures.

DR. BRATZLER: Patrick, this is Dale. So again the denominator I think eventually comes up here, too. In your composite, if you have a measure with a large denominator, it will definitely have a bigger
impact on overall performance of the composite, depending on the methodology and particularly if you don't weight it.

So certain performance measures, you know, the composites may look like they are performing relatively well if you have a large denominator for one of the measures that has high performance and other measures that may have smaller denominators with much lower levels of performance.

So without weighting, the denominator may affect the validity of the measure.

DR. KAPLAN: Can I ask a question?

The validity -- the purpose of the composite measure estimates some collective that is not represented better by individual components. So by definition you are measuring something different or somewhat different. So validity answers the question are you measuring what you think you are measuring.

So in that sense, the evidence
that is referred to in 1c isn't the right evidence. It is evidence at the patient level but it is not evidence at this level. So when you talk about validity testing, I think I would appreciate some clarification about what you mean by are you measuring what you think you are measuring if now you are creating some new collective of things that together are something else.

DR. ROMANO: Exactly. I think that is the point. I mean the evidence that we were talking about in 1c again is about the evidence about the components, not the evidence about the overall composite.

So this is where we have to demonstrate that, speaking as a developer now, that we are actually measuring what we claim we are measuring. And how do we do that?

These forms are difficult to fill out and when you get into these individual components, they don't seem to pertain, necessarily, to composites. I mean exclusions
what does that mean in the context of a composite because every measure that is part of that composite will have its own exclusions. But it is not -- that doesn't tell us about the performance of the composite, the validity of the composite as a whole.

DR. BURSTIN: Although we have seen exclusions at the composite level as well. So not at the individual component level but actually only at the composite. Patients who never make it into the bundle, for example. The resuscitation over sepsis as an example.

DR. BIRKMEYER: Well I think this criterion is going to be easier for some composite measure issuers than others. Those that have composite measures that are derived against some empirical standard and we talked about this before the leapfrog survival predictor, you could easily assess the extent to which that measure does better or more
poorly against like other measures of mortality.

If you took the composite measure for CABG by STS, which is basically kind of a four-part equally-weighted piece of mortality and processes of care and a few other things that are all kind of measuring different things, you would have no way of judging whether it measures what you think it is measuring empirically by the same way that you would in other context.

DR. ROMANO: So in that case, the validity would be intrinsically based on the validity of the individual components and the conceptual framework that they all belong together.

DR. BIRKMEYER: Face validity, too.

DR. ROMANO: Face validity.

DR. KAPLAN: But so for example, I am trying to estimate physician performance and I want to be able to attribute whatever
care is being provided to an individual. I could say well I am only going to apply this measure to people who the doctor has seen at least twice in the last calendar year because otherwise, I am attributing this care to a provider when that is really -- that isn't the primary provider of this patient's care.

So from the exclusion standpoint, that might be a very reasonable thing to do. But then I am still stuck with am I measuring physician performance? Am I measuring what I think I am measuring? And how are you going to tell?

And so doctors who provide good diabetes quality should do what? You know, should provide other kinds of quality, have lower overall something rates? What should doctors -- so that is how you tell if you are measuring what you think you are measuring. You either get construct validity -- you have no criterion validity, so you can't use that. But at least you should have some idea and if
you haven't already tested it, a direction, I would think, Helen, might be a direction to go, at least point us in a direction. If you don't have good evidence now, at least tell something about how you are going to evaluate what you think you are measuring going forward.

And people who are developing measures should be able to tell you at least something along those lines.

DR. DE LONG: And that relates to the comment about the STS measure. Because if it truly is a valid measure, then as time goes on complications, the individual complications should go down. If it is used for quality improvement, mortality should go down. That measure should be going up. And those should correlate as time goes on.

DR. BIRKMEYER: But those comments are no more true of the composite as they would be applied to the components.

DR. DE LONG: Well, it includes
the process measures. And if we are driving
up those process measures, are we seeing the
whole profile improve?

DR. BURSTIN: This is again an
issue for us about individual measures as well
as well as composites. It is often hard to
figure out what the gold standard is against
which to compare to know that you have got a
valid indicator.

So I don't know that I see
anything unique and different about
composites, beyond what is written here.

DR. ROMANO: Yes, I think what we
heard is that there may not be criterion
validity, for example, because we are talking
about a measure -- like let's say we are
talking about outcome measures.

So if we are talking about an
outcome measure of a particular type of post-
operative complication, then we can present
evidence of criterion validity based on some
gold standard of medical record review or
whatever. But if we are then putting together a bunch of those measures into some kind of a composite measure of patient outcomes, then either we have to fall back on the individual components and say that the individual components had criterion validity and therefore the composite does or we have to use a different validation framework and say well this composite is valid because it predicts the future outcomes of the patients.

And so I have some other evidence that I am going to use to show that this in fact predicts which hospitals or which doctors will provide better care in the future or better long-term outcomes.

DR. KAPLAN: Yes, see I think it is different from the individual component measures because the evidence base is attributable back to patients and what happens to patients over time. But those measures aren't necessarily a reflection of physician performance or an individual physician's
performance. It could be a collective of physicians' performance but I think the evidentiary base for using it now to reflect physician performance as opposed to good health outcomes for a patient, is a different -- that is a different measurement task and it needs a different kind of support.

DR. BURSTIN: I think I could probably have the exact same argument about some individual level measures as well. I am just trying to keep us on task. I agree completely those are really important conceptual issues. I just don't know that they are any different for a complex individual measure versus a composite. We have just as many issues with those kinds of things for an individual intermediate outcome for most docs, too or clinicians at all.

DR. ROMANO: Well I guess the question is just that developers should be asked to clarify whether their evidence of validity comes from the validity of the
individual components or whether they are making some broader argument that is based on the validity of the composite and testing the validity of the composite through construct validity or possibly criterion validity or something else.

At the end of the day, steering committees can decide what is acceptable and what is not acceptable but I think the idea would be just to force that decision point. Because if the developer is saying I don't have any evidence about the validity of this overall concept, aside from face validity, it makes sense to put all these things together, then it forces people to look at the individual components and to put more attention to whether the individual components are valid. And at the end of the day, they may decide that despite that, they are still not totally convinced that the composite is valid.

DR. BURSTIN: So does that go back
to the point we were talking about earlier, perhaps that you would give, it would be acceptable to have validity of the individual components but you might get higher points if you actually have validity of individual components and validity of the composite?

Would that be a higher level of a pass on validity, for example?

What if you only had validity of the composite and not the -- that was the third one. Sorry, I couldn't help myself. You are on a roll. What if you only had validity of the composite but not validity of the individual components?

DR. CHASE: Well again, I think that is only a problem when somebody is questioning whether all the components are necessary. I mean if you sort of already come to I think all of the components are necessary and then I am testing the validity of them all, that doesn't seem to be -- Great. You don't need to test the individual if you could
I actually do it. I think the problem is generally you don't have, you haven't tested them altogether. You haven't been able to do that. And so most often I would guess you get people who would bring a composite in with the validity around each one.

But I don't think we should bind this to say oh, if you can't prove both, it doesn't work.

DR. ZASLAVSKY: If someone handed me an example like this, I would scratch my head and go back and once we really look closely about why I didn't think that the components were valid but the culmination was because maybe my events for the validity of the composite has to be examined more closely.

You know, you can have a regression model that is predictive using a bunch of really inane valuables and then you figure out it is really because they are measuring the quality of the reporting or something else like that as irrelevant.
So I don't know. We don't need to belabor this. I don't think it is going to go into the criteria but it is certainly that something in practice that you probably look at more closely when you can't understand why the composite is valid but there is some evidence, there is some empirical nature.

DR. KAPLAN: That is like the betas are significant but the model is not. You end up with the individual components having significant beta coefficients and the whole model is not significant. So yes, individual components can contribute to something but it is meaningless.

So I think if the model is significant and the individual components are not, then you really are in trouble.

DR. ROMANO: There is this concept is described in the briefing memo to a balancing measures within a composite. So for example if you are concerned with readmissions that by focusing on readmissions you are going
to basically encourage hospitals to keep
patients longer in the hospital and basically
never discharge patients so they don't have to
worry about readmitting them, then on the
other hand, other people are measuring length
of stay and putting all the focus on
efficiency and get the patients out. And then
who cares if they get readmitted?

So by putting the two measures
together, you could argue that they are
balancing each other's weaknesses and leading
to a more valid composite measure than either
of the components alone.

DR. BURSTIN: I mean a specific
measure, this was when I first came to NQF,
was a measure Leapfrog had actually put
forward that looked at it was actually a
length of stay measure. And what they put in
am a balancing measure, which I think
ultimately got redone and isn't this way
anymore, but just for the sake of argument,
they included a seven-day readmission measure
in it as part of it to actually show that if
you are pushing down on length of stay, are
you actually going to then see it bubble up
with early readmissions. I think that is the
logic of it. And in fact, there is a lot of
concern these days as we have moved to sort of
bundle payments and lots of other purchase.
Do they need to be measures that sort of get
a stenting or potentially these balancing
types of measures. That is an interesting
argument of why you might have measures within
a composite that aren't going to -- would
never really work as a stand-alone.

DR. ROMANO: It would be a clear
conceptual basis.

DR. BURSTIN: Yes.

DR. ROMANO: Okay, so other
concepts on validity? Okay. So I think --
what is next?

DR. BURSTIN: Can I just clarify
what you are saying what you are saying so I
understand? So we are saying that you could
validity of the individual components or you could have validity of the composite either. And that if you do both, that is like gravy. That is even better. Yes? Okay, just checking.

DR. BIRKMEYER: But you can't have neither.

DR. BURSTIN: You cannot have neither.

DR. ROMANO: You cannot have neither and it may be sufficient -- I mean when you say it is better to have both, it may be sufficient to have one without the other, particularly to have validity at the composite level without demonstrated validity of all the individual components.

DR. KAPLAN: It is really hard for me to do this in the abstract because -- I said that backwards, by the way Alan, it is like having a significant model with no significant data. Sorry. But you know trying to think through, Helen, what would be an
example where for the purpose of assessing something new, creating a collective out of that that had no accuracy individually that now you are going to summarize and make into now something that has accuracy for some new purpose.

You know I am still struggling with how that would work.

DR. BIRKMEYER: This isn't my particular field of expertise but sort of the implementation side of this can point to lots of illustration where bundles of processes of care, if they are all done together lead to salutary effects whereby all of the evidence shows very negligible effects of any one thing, like UTIs or SSIs after surgery.

DR. ROMANO: And I think the more common scenario even is where you really are unable to get evidence about the individual -- the validity of individual components because if it is a rare event and it is just not feasible to assess the criteria and validity
of some of those components.

DR. KAPLAN: Right. Again, it comes back -- I don't want to get tangled up in evidence because this is a very theoretical discussion. But it is about are you measuring what you think you are measuring?

So adding those things up in quality terms, the evidence support comes from a different purpose. Now I am creating a new purpose of measurement. Now I am going to create a physician performance measure out of this. What is the evidence that that is what you are measuring is physician performance, not patient outcomes or patient something or other.

So that is where I still am struggling with the composite versus the individual components when it leads back to the evidence that NQF is required for all the individual components.

DR. ROMANO: 2f in the current measure evaluation criteria for composite is
about methods scoring and analysis that allow for identification of statistically significant and practically or clinically meaningful differences in performance.

Any questions or concerns about that? So it is basically the same criteria -- the same criteria exists for individual measures. So that seems fairly straightforward.

Disparities, 2h again is the same criteria I think as for individual measures.

DR. BURSTIN: Yes.

DR. ROMANO: Right? Okay. So what is different for composites in the current framework is 2i, j, k, and l. So let's focus our remaining discussion on those items, 2i, j, k, and l. And this is where people have a little trouble.

So 2i is about that components -- empirical analysis showing that components fit the conceptual construct; 2j is about contributing to the variation and the overall
composite score; 2k is about weighting rules that are consistent with the conceptual construct; and 2l is about how missing data are handled.

DR. SHAHIAN: I think those are pretty good.

DR. ZASLAVSKY: I'm a little uncomfortable with 2i because I think if you are doing composites you are probably moving to or likely to be moving in a direction of a broader kind of construct for which the internal consistency is not going to be as high.

I think there is certainly circumstances where you might look at it but it is only one of the possible arguments for creating a composite.

DR. SHAHIAN: Would the or statement take care of that?

DR. ZASLAVSKY: Well I would rather not see one thing highlighted there and then other things being something you would
have to justify.

DR. ROMANO: Yes, I would agree with that. Personally, I am comfortable with the assumption that internal consistency reliability has to be met or if not, there needs to be justification because, as we have discussed earlier in some cases internal consistency reliability is just not appropriate for the purpose of the composite.

So how do we reframe this in a way that it is more inclusive in terms of linking the methods that are presented with the purpose of the composite?

DR. DE LONG: Can we just take out what is in parentheses?

DR. ZASLAVSKY: Yes, there might be some -- I don't know if there is some explanatory material maybe a footnote that would refer to some of the different types of analysis that would be relevant to different purposes but I wouldn't put it into the item itself.
DR. ROMANO: So I guess looking together at 2i and j, I think my concern would be that these two items put too much emphasis on the components and not enough emphasis on the overall properties of the composite. So is there a way to shift that methodologic focus a bit?

DR. ZASLAVSKY: I think i and j were written with a view to particularly the value of a composite for improving reliability. And since we are looking at a broader set of purposes, that might be one thing you might look at that would be one reason for the composite, one support for the composite but it is not other rationales which would not involve that. I'm not quite sure how to deal with that. It sort of makes that a default the way it is written now.

Maybe the point should be more a rationale should be given for inclusion for all of the items or something like that. We had a bit of a discussion about that before
about whether you would want to have a bias in favor of more parsimonious composites when there isn't an argument for including everything just to make them easier to create and reduce the data collection burden. But is a very qualified argument. So it is a little hard to formulate it as a general criterion.

DR. ROMANO: So is there a way of taking 2i and 2 j and reframing them so that it is linked a bit more like an if-then sort of logic? Like if the developer says that the purpose is to increase reliability, then we look for evidence related to internal consistency reliability.

Helen doesn't like that idea.

DR. BURSTIN: It doesn't lead to consistency in our committees and it leaves it up too much to the developer to make that call.

DR. ROMANO: Well I am trying to come up with a way to improve consistency so that it would be clear that if the developer
says A, then the committee is expecting to see B. If the developer says C, the committee is expecting to see D. So I am actually trying to see if that can be improved by tightening the linkage between the measurement construct and the evidence, the validity testing evidence that is presented in support of that.

DR. ZASLAVSKY: Patrick, here is another tact for 2j. What if we asked that the application shows reasonable attention to parsimony as a value? If that is the reason for this, let's just state it as a value directly and then the developer can respond by saying this is why all of these items are important or they can get that one and say you know I could have dropped out five things and it would be just as good. Is that what we are trying to get at there?

DR. ROMANO: Feedback on Alan's idea?

DR. KAPLAN: The way it is worded now you can't do it with an index anyway
because each item has a zero/one probability
to contributing to the variants in the
outcome. So you couldn't even use this for
indices.

So for me, if the intent is that
you want to have items that improve precision
to the level you are shooting for for the
purpose you are trying to put it to and the
addition of items beyond that that don't
contribute unique variation if there is unique
variation to contribute to the overall score,
then your rationale for including them has to be something other than improvements in
precision.

I didn't know the way that reads
if that was your intent. What was the purpose
of writing 2i and j, Helen?

DR. BURSTIN: I'm not sure I
remember exactly but I do think part of this
was because everything above it, 2a through 2f
was all about the composite itself at a higher
level. I think this really reflected the
concerns about the component measures. So I think this was again a look -- this was why I think we have had that discussion is about composite level of the individual score level or both. And I think the composite committee last time squarely came down the side of both. And so I think the question for me is are we still in that same place? I'm not sure we are. But I think it was to say can you also justify the individual's inclusion in an empiric way?

DR. ROMANO: So Sherrie, what would you propose instead of 2i and 2j? I mean I completely agree with what you said. I am just not sure how to frame it in these evaluation criteria.

DR. KAPLAN: I think I get the principle that they should, that the components of the composite should share something in common. I think that is what you are shooting for empirically. Right?

DR. ROMANO: They should tell us
something about the construct.

DR. KAPLAN: That they should share something in common that reflects an underlying construct, the latent construct, whatever it is. All of the things it used to represent that should at least share something in common.

DR. ROMANO: Well they don't actually have to share any variants, do they?

DR. KAPLAN: No.

DR. ROMANO: Right. Okay.

DR. KAPLAN: No, not at all. Because if you do an index, by definition you are going to limit your ability. And for things like a collective measure of comorbidity, it doesn't make any sense to do that kind of analysis.

DR. ROMANO: So when you are talking about sharing something, you are talking about sharing something at the conceptual level.

DR. KAPLAN: Well either -- yes,
either the conceptual -- they have to have one
of three things. One is first they have to
have some kind of conceptual nod from the
people who are experts. They had better share
that in common. You are measuring what you
think are measuring or you are doing -- your
measurement error isn't going to be too great
so that you are actually reflecting physician
performance or hospital performance or
something, one.

    The second thing is is I would ask
for some kind of either plan for or supporting
evidence for the composite different from the
individual elements of it for the reliability.
And then I would ask for some evidence that
individual components contribute uniquely to
the overall construct, that you are not just
adding more things because you can measure 50
things about somebody's quality of care but
you chose the eight things that really are
sensitive indicators to good quality.

    DR. ROMANO: Okay so that gets
back to Alan's idea about parsimony, that there has been attention to parsimony and the construction of the composite and that therefore the components of the composite have to have some justification either an empirical justification or a conceptual justification.

If we are relying entirely on a conceptual justification, presumably it should be a fairly strong concept. Is that fair?

DR. DE LONG: So how does that take us away from what is written? It seems that that is consistent with what we have here.

DR. ROMANO: Well it basically erases $2i$ and $2j$ as they are currently written and rewrites it in a more general context.

DR. ZASLAVSKY: I think the concept of $2i$ is still pretty much there except we have taken out the specific of the item. But I think $2j$ is a very specific notion of why you would include items and we are just broadening it. We are saying you
just have to had paid attention to having a rationale for including the items. It doesn't have to be about reliability.

DR. ROMANO: I think in general where we are agreed is that 2i and 2j are too embedded, as they are currently written are to embedded in a particular framework or approach and that we have to come up with much more general language that encompasses a variety of different applications. The devil will be in the details of the wording here.

DR. BURSTIN: And actually just to pull from the old report because you asked why that was, so they specifically had in this section on scientific acceptability, several approaches might be used to combine measures. One approach might be the psychometric approach developed, blah, blah, blah, create a complex construct that is not directly measurable using multi-item scales.

With the psychometric approach, the component items are measures that
generally are measuring the same underlying construct and should be correlated with one another, although not perfectly and they would be redundant. Some composite measures may not reflect this classic psychometric construct, depending on the types of items or measures that are included in the composite. When the components are not correlated, the rationale and justification for their inclusion must be provided and appropriate analyses identified.

So that matches our discussion. We will just have to see how that translates. That is actually, from this discussion they have got this wording. So we will need to see how those play together.

DR. KAPLAN: Just to follow up on that Helen, so for example, if you are looking at physician performance and sort of how you get to reliability at the physician level. So one common way to do that is interclass correlation. So you look at how much do
doctors, are they consistent across patients
in their practice and do they differ from
other doctors in your comparison group?

So that is a different approach to
sort of the same principle, the issues of
principle. But this sounds to me like I did
a total correlation matrix and that doesn't
make any sense for a lot purposes.

DR. ROMANO: It is also a little
bit confusing about whether it is a
reliability concept or a validity concept,
frankly. Because I would think of item total
correlations and Cronbach's alpha as being
reliability measures, not as being validity
measures but they are here under validity.

DR. KAPLAN: But you could an
exploratory factor analysis, for example, and
then you are sounding very similar and that is
often used as an evidence for validity.

DR. ZASLAVSKY: Or you might do
something like a criterion regression like
regression of mortality on a bunch of
variables that would say these are all contributors to mortality.

DR. ROMANO: I think back to the work that John did with Leapfrog. Basically and correct me if I get this wrong but I thought it was very clever because the idea was that consumers and people acting on behalf of consumers, employers, purchasers, whatever, want to pick hospitals based on where is the best place to go today for esophageal surgery.

But what we have are these data from two or three years ago. And so what we really want is to figure out a way to bring together multiple measures to get the best estimate of current performance. And that may involve compositing several different types of measures in order to provide the best prediction, if you will, of current performance to inform current decision-making.

So this gets back to my fixation on decision-making but if the decision-making framework is about helping consumers and
purchasers make decisions today about where is
the best hospital to go, then you can develop
a conceptual rationale and you can test the
validity based on how those components
contribute to a better prediction, a less
biased prediction.

DR. BIRKMEYER: To be fair, I
think we were also a little bit unique using
the tool of seeing how well historical
measures forecast outcomes in future years, as
a twist on the usual splits sample approach
but again trying to get back to what you are
trying to simulate or what the consumers are
actually using the data for and it is to make
a decision here and now.

DR. ROMANO: Okay, so let's look
at 2k, the scoring/aggregation and weighting
rules are consistent with the conceptual
construct. And then there is some stuff in
parentheses that I would disagree with. I'm
not sure how others feel.

But I guess the question is do we
want to indicate a preference for a particular weighting scheme? I would argue in general that any weighting scheme involves value judgments and, therefore, equal weighting entails a particular set of judgments that doesn't make it any better than anything else.

DR. SHAHIAN: I would just say that the justification for any weighting scheme or lack thereof must be given.

DR. ROMANO: And what guidance will we give to steering committees to evaluate that weighting scheme?

DR. SHAHIAN: I think you have to ask if they do choose weighting how are the weights derived from factor analysis, from -- you know we actually in the STS we tried to figure out among the various morbidities what is most important to patients. We tried to figure out what providers who had seen many patients with strokes versus death versus internal infection, how they graded the relative importance of those complications.
Let me just tell you that in cardiac surgery it is very hard to do that but I'm sure in other areas people have done that, sort of quality of life impact and that sort of thing, expert opinion, delta, whatever. I think one just has to use one of those methods.

DR. ROMANO: Well I like that because it puts the emphasis on kind of a patient-centered weighting scheme of what is most important to patients, what has the most impact on patients. And I really like that. It would be nice to encourage more of that sort of thing.

DR. BIRKMEYER: Well we just to need to acknowledge that in the short-term that a lot of those judgments are going to be based on expert opinion.

DR. KAPLAN: Just as an experiential cautionary note, anytime you mess with any sort of otherwise sort of straightforward activity like adding things up and you move it further and further way from
common sense, you get harder and harder to
explain to people and it is less and less
transparent.

So I would like to make the case
for at least they better bloody well compare
it to what happened with a simple algebraic
sum or some simple add them up of criteria
that is such an improvement that it shuffles
the whole distribution around. It does
something else that is a substantial
improvement over what would otherwise be a
little very much easier to translate and
transmit measurement activity.

DR. ROMANO: You are favoring
equal weighting as a default.

DR. KAPLAN: I'm favoring a
summary of things that is easy to transmit and
on the face of it looks like not a
transformation that is going to introduce
hocus-pocus.

DR. BRATZLER: It almost comes
down to whether you weight them equally or
with regard to weight composite measures, the individual components about whether you should even include some of your components. I mean you almost to need to make some argument if something has to be de-rated, perhaps it would need to be in the composite.

DR. DE LONG: I think it is not altogether clear when we talk about equal weighting because, for example, some components will have fewer observations involved. So do we transform them and then add them equally or do we add them in equally de novo?

So I think we are always going to get into some complexities that just need to be explained and rationalized.

DR. ZASLAVSKY: I'm going to suggest something parallel to what I suggested on the inclusion of items that you just expressed the principle of parsimony that here we should just say with due regard to simplicity and presentability as well as other
justifications, something like that. Just express what the value is and then the committee is going to know they are supposed to look for that and see whether there is a rationale if things are very complicated.

DR. DE LONG: I just wanted to say they should recognize that the more complicated they make it, the harder it is to get evaluated. So they are probably going to want to make it as simple as possible anyway.

DR. ROMANO: I mean I guess I sort of agree with what you said earlier that even -- there is no avoiding complexity. And that even things that seem simple aren't really simple because they involved certain assumptions that are pretty wild assumptions.

So when I see process measure composites that just add together ten processes and the denominator is five times bigger for one of those than another one and another one has five times the impact on death as another one, I just wonder what the hell.
I mean, it doesn't have face validity for me to add those up because they are so different in terms of their impact on patient outcomes on what really matters.

DR. SHAHIAN: Yes, and just to support what Sherrie said about messing with simplicity, after trying all the things that I mentioned in the STS we ultimately defaulted to equal waiting of the various components. And in fact most of the literature that I could find at the time including what I think is the single best reference on devising composite measures, at least for healthcare is this OECD document by Nardo which I think is great. But they all suggest equal weighting. So I think we have some justification for that.

So I think just providing a justification for the approach that you do and whether we acknowledge that sometimes simpler is better.

DR. GOLDSTEIN: And I guess I
would second some of the things that people have said that just really providing a justification for the weighting scheme, I mean for your health plan ratings for example, we got actually lots of complaints from stakeholders when we when we did equal weighting. So we did spend a lot of time getting input from clients and consumer advocates and all different experts on alternative weighting schemes. And then in the end we talk about input, and it was a policy call as well, and incorporated weighting. And that is probably one of the things that we did at least for that rating system that we got the least amount of complaints or concern about because people generally agreed we weighted outcomes the highest, process measures the least, patient experience kind of falls in the middle.

So where people quibbled is the size of the weights but I think in general we moved the industry in the right direction. I
mean everyone agreed all stakeholder outcomes ultimately is where you want to go. So that should, in any composite measure for combining different types of measures should be weighted the highest.

So I think in any submission you really want to understand the rationale and what are you trying to do with the measurement. What are you trying to drive in terms of quality improvement. What is the most important indicators? Maybe those are the ones that get the highest weight.

DR. BURSTIN: How does all-or-none fit in here or does it?

DR. ROMANO: Well presumably it would require a rationale that is consistent with the conceptual construct. Right?

DR. BURSTIN: Right.

DR. ROMANO: So the developer would have to present a conceptual construct in which all or none weighting would make sense.
DR. BURSTIN: Right. Well I'm just wondering if we need the first sentence at all or do we just being the parentheses with weights are determined by empirical analysis or a systematic assessment of expert opinion. I mean it just seems like I don't know that we need to state that equal waiting is preferred.

DR. ROMANO: Right. I think that is -- I am certainly agreeing with that. I mean there is some -- in place of that wording there is some suggestion of again perhaps more general wording saying that in general simplicity is good. That doesn't necessarily mean because even in the concept of equal weighting is not entirely clear, as Liz pointed out is what does that mean in terms of how an indicator is actually constructed. Is it the denominators are equal? Is it the numerators are equal? Is it the standard errors that are equal?

DR. KAPLAN: Just empirically, the
most robust scores are the ones that are robust across weighting schemes. So you get the same answer no matter what you do. So in some ways this falls back to the measure and how they are constructed, rather than ultimately how they are weighted or not weighted for a scoring scheme.

DR. BRATZLER: Well and again, it is part of the reason we have never submitted any of the composite all-or-none measures that we have developed in the past and have used them for performance improvement but not for a calibraity. If you have one measure that particularly in the classic all-or-none calculation, if you will then measure it with a very large denominator, when your composite really ends up simply reflecting to a large extent performance on that individual measure and not your other measures that you may have as part of this composite. So you do have to worry about denominators in all-or-none measures and if they are all rated equally,
then the composite may just reflect what is in
the measure.

DR. ROMANO: At least in the AHRQ
composites we have incorporated reliability
weighting so that the measures that have more
hospital level signal get more weight based on
the constructs that we are trying to inform
the market about hospital performance in a
particular domain. So it would make sense
that measures that have more signal at the
hospital level would better inform that
overall decision but might not be right for
all cases.

DR. BRATZLER: Well that certainly
would be the approach that CDC is looking at
for some of the infection measures now, which
is reliability rated.

DR. ROMANO: Okay. So any other
comments about scoring and weighting rules,
what developers should be told there? We can
talk about 2l, which is about missing
component scores. Do we want to revise that,
delete that, add to it, clarify it?

DR. SHAHIAN: I wonder if we should say anything specifically about the management of missing scores in all-or-none measures. There I think it becomes there really missing data on one component of an all-or-none can affect the all-or-none measure in a way that it doesn't affect other sorts of measures. I don't know how others feel about that.

DR. ROMANO: Well, the same thing is true for validity of components. Right? So I mean in an all-or-none construction, the validity of one component may drive the validity of the entire composite in a way that wouldn't happen with other weighting schemes.

DR. ZASLAVSKY: I think I agree with the content of 21 but I can't parse the sentence. Am I the only one who finds this sentence hard to read?

Analysis of missing component scores supports the specifications for
scoring/aggregation and handling.

Isn't there something about what we want that to actually show that it will eliminate bias associated with -- will minimize bias associated with missing data, something like that.

DR. ROMANO: I like that emphasis on minimizing bias.

DR. KAPLAN: And specification. They have to say what their missing data, how the missing data treatment, what treatment of missing data is and then how that treatment minimizes bias. I would stick with some language with that. I agree with Alan, that is a very complicated sentence.

DR. ZASLAVSKY: And they should tell us how much missing data there was in their pilot data as well. If that is really high, you have to really question the feasibility.

DR. CHASE: And the only thing I would caution about this would be not having
a preference for missing data isn't a reason
to score people low. I think many in the
measurement feel for example smoking status,
if you didn't take it, it is alright to count
it as not having -- you know you get a zero
for that.

So to me that isn't a measurement
error, as opposed to other things where data
just wasn't available or you didn't pick it up
in certain places where you should have. And
then how you deal with that I think is
important, especially because a lot of these
composites, as you know, would take the mean
of the whole or something and that adds extra
change to the measure that may not be fair to
everybody.

DR. KAPLAN: Helen, are you going
to use this? Are you going to score this?
Like in the Cochrane stuff, if you didn't put
anything in about how you treated missing data
you get a zero. If you put some lame thing in
that isn't very good but at least you told
people, you get a one. And then if you did a really super-duper job of it you get a two.

So are you going to score this?

DR. BURSTIN: They will get rated overall on the score for validity and the score for reliability. So this will factor into it, yes.

And just one last thing in here, I think part of the reason this was also here is some of the composites --

DR. ROMANO: It is not necessarily a point scheme. Right?

DR. BURSTIN: No, it is not a point a scheme but it is factored in. So if is left out, we would send it back to the developer to finish. It has got to be complete. Yes.

But just one other point about this. I think one of the reasons this was here is we had seen some composites, for example, that had assigned the mean national value for example for missing data which
committees just didn't think were kosher and
make sure that is okay that that is captured.

          DR. ROMANO: There are some
variance problems there.

          So what I am hearing maybe for
both 2k and 2l is that we actually want more
specificity. In other words, we are maybe not
going to be prescriptive about what developers
must do but we are going to demand more in
terms of explanation and justification.

          So in the case of missing data,
they should show us how much missing data
there is on each component and what their
approach was to handling the missing data.

          As Jim pointed out, a perfectly
acceptable approach might be to assume it
didn't happen for the sake of some measures
but that should be explicit.

          In other cases, they may have done
some imputation but if that did some
imputation, it better be an approach that
incorporates some variants and just assume the
same value.

So we will expect that to be submitted as part of the process. And perhaps the same for 2k that developers have to show that their weighting scheme is consistent with the conceptual construct. And to the extent that they might have compared their scheme with a simpler scheme, they should share that finding.

So other thoughts about these components? I think is it time for us to take our afternoon break probably? Okay.

All right, well let's take a ten-minute break. We will reconvene at 3:15.

(Whereupon, the above-entitled matter went off the record at 3:01 p.m. and went back on the record at 3:16 p.m.)

DR. ROMANO: Okay, so we are reconvening. Thank you, Dale, for staying with us.

DR. BRATZLER: I am here.
DR. ROMANO: Okay, so now we are moving on to talk about NQF measure evaluation criteria three and four; three is usability, four is feasibility. And here we are actually doing a little last minute work because the usability criteria are about to change in a fairly substantial way.

DR. BURSTIN: It is correct on 12. Page 12 is correct.

DR. ROMANO: Yes, right, except it is so cryptic there that I asked them to bring it up the full version.

So what is now called usability will be called usability and use. And if you look at page 12 of this briefing document, the DRAFT Table 4, it lists very cryptically three criteria for usability and use but we are trying to pull up a more specific version so then we can see how these would be modified for composites.

(Pause.)

DR. ROMANO: Okay. Anyway, sorry
for the technical delay. But the two additional criteria you can see 3d and 3e. We can start talking about that a little bit. So 3d is saying that data detail is maintained such that the composite measure can be decomposed into its components to facilitate transparency and understanding.

So let's --

MS. PAGET: Patrick?

DR. ROMANO: Yes.

MS. PAGET: Just to add a little light to the afternoon. If you do a thesaurus check on decompose, you get rot, decay, crumble, fester, putrefy.

(Laughter.)

MS. PAGET: So if we do want to stay away from that word, I think we could simply say data detail is maintained such that the components of the composite measure are transparent. I mean, you know, --

DR. KAPLAN: We could use disaggregated.
MS. PAGET:  It has to have a fancy word.

DR. ZASLAVSKY:  I am going to look up the thesaurus definition of transparent and you are going to pull that back.  Can be seen through, invisible.

MS. PAGET:  Try unpack, unravel, and deconstruct.

DR. ZASLAVSKY:  How about disaggregate?

DR. ROMANO:  Well I think the conceptual problem is that we have already talked about some scenarios earlier today where in fact a composite is being constructed of components that do not support disaggregation, at least for public reporting applications.

So is that okay?  I mean disaggregation may be desirable to providers to see where they went wrong, so to speak, within the composite.  But if the component completely randomly distributed, then it
doesn't actually inform public reporting.

DR. KAPLAN: I was waiting to see if Helen turned my microphone off.

So to the extent that people are going to want to know how to improve their scores, they are going to want to know where did I fall down. So these things, if you kind of report them back out in any kind of disaggregated form, I think people are going to be bummed out. That is a technical term.

DR. ROMANO: Right. So does the disaggregation have to be subject to public reporting?

DR. BURSTIN: I think the question is it could certainly be used for internal QI. I mean people use all kinds of back of the envelope stuff for internal QI. I guess the question is, for example, if you have several component QIs in some of your components that you don't feel are reliable, estimates on their own, you would certainly would not want to use those for accountability. So I think
there probably needs to be some statement that they should really only be used in that way if they are indicated as being reliable at the individual level. Beyond QI only. Does that make sense?

DR. ZASLAVSKY: Well they might not be -- yes, I guess for public reporting, yes, or for payment. For QI any level of detail is fine.

But even there in some cases the components aren't meaningful at all. You know like if it is an A or B type of thing. So you don't want to impose too high a standard there breaking down something that isn't really meant to be meaningful in its pieces.

MS. PAGET: Well there is transparency and then there is intended uses for purposes of -- so could the developer be asked -- I mean at minimum we want transparency, I would hope, just to be able to define what is making up the composite. And then perhaps secondarily they can indicate its
uses for other purposes. I don't know.

DR. ROMANO: Yes, I guess what does this mean in practice? Data detail is maintained such that the composite measure can be disaggregated. What does that actually mean?

DR. BURSTIN: I think the intent was simply that for putting forward a composite measure there should still be the capacity to be able to look under the hood and look at the five component measure scores that went into it.

This does seem like pretty highfalutin language to just say that. And maybe it is just something as simple as saying that when appropriate, based on measure score performance -- not measure score performance but performance of the measure characteristics being able to examine the individual components should be encouraged for something along those lines.

MS. PAGET: Well and if we really
think that we are someday using these composites for value-based purchasing and comparative reporting, et cetera, et cetera, you have to be able to look under the hood, I would think.

DR. ROMANO: So would anybody ever fail on that criterion? I mean is that just something that is expected or --

DR. BURSTIN: I mean well I guess that is the question. I mean not to pick on the AHRQ example but if you have a composite that you feel very comfortable has great performance at the composite level but you don't have great comfort, necessarily, in some of the performance at the individual component level, would you always want those to be reported out such that -- you know we just talked about this earlier. Would you want a consumer looking at a non-reliable indicator? Is that what we want to do?

So I guess for me that is the one qualification is you should really only put
out there what you feel comfortable is in fact a valid representation of quality to consumers and purchasers.

DR. SHAHIAN: I think the providers, though, and we certainly have seen situations with the STS CABG composite where the providers really wanted to know or challenge how a particular domain score was arrived at and we can provide that to them now routinely.

So I think at the very least providers have to have the ability to get to the detail level. Some may not be appropriate for public reporting but at the very least the providers that are being judged by these measures need that.

DR. ROMANO: So then what I am hearing is the emphasis really is on the last part of this. It is on facilitating transparency and understanding. And somehow we need to rewrite the first part of this to make it clearer that we are not -- that what
we want to ensure is that the data are
collected and composited in a way that permits
this disaggregation. But we are not
necessarily asking developers to support that
disaggregation for every application of the
measure.

DR. SHAHIAN: I think the ideal
scenario is one in which a less-informed
consumer can look at something simple and
visual like a star, a more sophisticated
consumer can drill down to the next level of
detail what is behind that rating. I don't
know if we want to express a preference for
that but multiple levels of detail are
available, something like that.

I think that various stakeholder
differ in how much information they can want
or digest.

DR. BURSTIN: We don't want to go
into that here. That is not without
controversy, as we have seen.

But I mean this is change from the
prior version. I did pull up the prior version. The idea over there was that it is critical that a composite measure when reported is readily decomposable into its constituent domains and individual measures. This will focus and facilitate quality improvement activities by providers and increased transparency and understanding of the measured results by all potential audiences. Additionally, it should be demonstrated the purpose of creating a composite measure was achieved.

So they are actually stronger. They are saying that -- and again part of this was because they required that the individual components had to essentially meet all criteria. So in that case of course you would put them out there. But if that is a potential change that this committee is recommending, then I think that does need to be qualified.

DR. ROMANO: Okay, so people could
I guess I spend too much time in the classroom.

So the current criteria or the new criteria here focus on -- no this is not the new criteria, are they?

(Pause.)

DR. ROMANO: There it is. There it is. Okay. So 4a is about -- so what I want people to do is just look at these because this is what is coming down the pike as of January, I guess.

So do these need to be adapted or modified for composite measures? So 4a is about accountability and transparency. And the idea here is to focus on use. In other words, usability is manifested by use. If an indicator is really usable, then it should be used. And so criterion 4a is putting forward a specific criterion about the actual use of a measure. Is there any reason why this criterion would not apply or would differ for
composites?

Not hearing any, let's look at 4b.

So 4b is about demonstrating improvement,
ultimately this is all about improvement,
achieving the goal of high quality efficient
healthcare for individuals or populations. So
new measures get a pass here but the idea is
that -- not a complete pass -- the idea is
that there should be a rationale for how the
performance results could be used to prove the
goal of high quality efficient healthcare for
individuals or populations. And if a measure
has been in use, then in fact that would be
demonstrated.

Any thoughts about how this
applies to composites?

DR. KAPLAN: Are we in 4c or still
on 4b?

DR. ROMANO: We are on 4b.

DR. CHASE: So it is interesting,
I don't see a difference in this for
composites versus not. But it is -- I only
question this now in looking at it in the sense of something like consumer satisfaction, which we might want to measure and people might find valuable and so forth. And even if didn't change, people might still want to look at that.

Now again, I am all into all the stuff we do we are always trying to improve so I hope there is change but I would hate to see us stop doing something only because we didn't improve it.

And what this sort of implies is an overall that after three or four years if you can't show any improvement in a measure then the measure goes away. Yay! We don't have to worry about that anymore. It didn't do any good.

DR. ROMANO: I mean at some point you would say that if you haven't caused any change that it is probably because you can't change that thing, even though you would like to be able to change it. You wish you could
change it. But gosh, people have tried for 20 years, they haven't been able to change it. There is probably a reason for that. But I agree that three years is too short a time horizon for something that is salient to consumers and patients.

DR. GOLDSTEIN: We actually for patient experience measures I think from the hospital side and health plan side, we have been seeing things big improvements.

DR. ROMANO: But there are certainly some outcome measures where we have not been able to seen an improvement and I think there is a reasonable argument to be made there that maybe those outcomes are actually much harder to improve than we think. Maybe we don't really know the mechanisms for how to improve them.

DR. GOLDSTEIN: I mean we are seeing for the Medicare program on the health plan side, we saw some measures that hadn't changed for years and years and years but once
you put the money all of a sudden they are
increasing a lot and there is more emphasis on
quality improvement. So maybe things they can
move if they have the rate incentives.

DR. ROMANO: So you shouldn't give
up until you have tried a wider range of
incentives.

Okay, so I think we are agreed 4a
and 4b are generally consistent for composites
but it really, I do want to emphasize for 4b
that it links back to the conceptual
framework, the measurement construct. So the
rationale that is described here has to be a
rationale that links back to that construct.

Okay, 4c is about unintended
negative consequences. So this criterion
emerged from a lot of discussion. I happen to
be on this committee but it emerged from a lot
of discussion about unintended consequences of
performance measurement in practice. And
consequences that may not manifest within a
single performance measure but maybe if you
had another measure you would see that
something else is deteriorating, that people
are gaming the system, whatever.

So this is now embedded from the
evaluative framework under usability and use.

Thoughts about how this applies to composites?

DR. BRATZLER: This is Dale. So I
think it definitely does apply to composites.

And one of the concerns that we had is we
started to working with composites such as the
all or none measures but when you had a group
of individual performance measures that
perhaps had substantial variation and
opportunity for improvement but as they
improved, when you look at composite weights
performance it made the opportunity to look
much greater with the composite measure. But
then the only way you could get very, very
high levels of performance on the composite
measure was to achieve near 100 percent target
on the individual measures and I think that is
problematic, particularly in the
accountability since there are not many performance measures that have 100 percent on the target, unless you have perfect specifications that will either be no exception to the performance measure. So by our experience with all-or-none composites was that it tended to drive you towards trying to achieve 100 performance on every single component of the composite, which could lead to unintended consequences.

DR. BURSTIN: Especially among measures that are otherwise at very high levels of performance. I don't think it is much of an issue when you are not at those high levels of performance.

DR. BRATZLER: That's true. And you know Helen, the one thing that we have seen in virtually every measurement we have pulled out in some accountability format has fairly rapid improvement in rates on the measures, you know versus the individual measures but they tend to improve fairly
quickly.

DR. KAPLAN: I got stuck on the word evidence because evidence of unintended consequences is a pretty strange concept, especially if you are going to get it by the journal editors who may want to love to publish bad things about quality assessment but evidence of negative consequences is -- and it may uniquely apply to composite measures because you are not sure what the collective of those things is going to do to you, especially in things like the all-or-none situation.

DR. ROMANO: Well so let's talk about that because again I happened to be part of that discussion. So I think there was concern that there is a lot of hand-waving about unintended consequences, where people say oh, terrible things could happen. This could happen, doctors could discriminate. They could send all the patients to Canada or Mexico. And you know at the end of the day,
it is all hand-waving because it is very hard
to demonstrate that in fact doctors and
hospitals stop being professionals and start
behaving like three-year-olds.

But so that was the goal here was
to say that there actually had to be
affirmative evidence of unintended negative
consequences. But maybe the wording is not
right. Is it different for composites than it
is for individual measures?

DR. KAPLAN: I'm just concerned
that there may not be the caliber of evidence
you wanted, evidence counting, capital
evidence as opposed to bellyaching.

But if there is evidence of actual
documented, published evidence that passes
peer review and all that scrutiny, different
from other -- I'm getting stuck because the
composite measures won't immediately be able
to tell you that.

Further, unless you disaggregate
you won't know if it was the fault of one of
the individual components of the composite measure or the fault of the collective taken as a group. I am listening to a lot of complaints about I won't even -- all cause admission.

But you know, there is a lot of complaining about what is going to happen, oh hand wringing and everything else with yet when we meet the evidentiary criteria here.

So for a composite measure I am a little bit more concerned that the evidence is never going to be there to the extent that you want it to be there. Because again it will depend on who you studied, whether that was just a squirrely group or whether it was generalizable to a bigger population, et cetera.

DR. SHAHIAN: Yes, I don't think there is any difference between individual and composite measures with regard to this particular concern. I must say, though, it is a very subjective assessment because there
are, for example, at least eight papers in
CABG and PCI literature over a span now of 15
years showing evidence of risk aversion
associated with public reporting.

Most people have assumed that the
benefits outweigh that risk aversion but I
don't know how you would ever make an
assessment of which is more important, the
advantages or disadvantages. But I think
there is evidence in some fields.

DR. BIRKMEYER: And that is
probably the only instance in which there is
published evidence --

DR. SHAHIAN: Yes.

DR. BIRKMEYER: -- about the
unintended consequences. And even then it
doesn't dissuade anybody's view about
measurement being the right thing to do. I
wouldn't object to scratching the whole thing.

DR. BURSTIN: Well these are
actually the general criteria that have been
approved for NQF for all measures. This isn't
specific to composites. We could certainly take a look at that question. But as Patrick pointed out, the real reason for putting evidence there was the concern that people would just put -- anybody could say oh I have real theoretical concerns about this measure going forward. And it is specifically supposed to be balanced to the benefits.

There have been published papers, I mean for example the ED antibiotic administration within four hours clearly published demonstrated. But you are right, they are few and far between.

DR. SHAHIAN: Premature activation at cath lab for BCI.

DR. BURSTIN: Right. So there are some examples. There are not very many. It is not as if it requires published Big E evidence to use your term. But I think there has to be some accumulating evidence out there that there is potential harm to patients and populations because of it.
DR. ROMANO: Okay well so relating this to composites though, it becomes trickier because there might be what if one component in the composite has such evidence but then there are nine other components of the composite that don't. So does that mean you should drop the one that does or does that mean that you should assume that the nine others counterbalance the one that does and so it all comes out in the wash.

DR. KAPLAN: Yes, that is very well said because that is what I was trying to get at. You don't -- with a composite, different from individual measures, you don't really know what caused the trouble unless you study it in a disaggregated way.

So you would have to build on the evidence that comes from the individual measures and make some assumptions about the composite or you would have to study the composite and then internal to the composite disaggregate it and look at the individual
what caused trouble.

DR. ZASLAWSKY: Does anybody --

can I have an example of a problem of this

sort that would arise with the composite that

would not exist within the component measures?

It seems like in general it has to just --

Sherrie suggested the composite would tend to

soften the impact of any -- let's say the

over-incentivizing of any one thing because it

is going to be mixing with a bunch of other

things.

So you would usually expect the

composite to be less likely to produce, as

anytime you average together a bunch of

things, it kind of generalizes and avoids

over-incentivizing some narrow piece.

So I'm not sure that we need to

state this as an issue for composites unless

the only kind of thing I could think of is

that someone would look at a composite that

combines several things and misinterpret it as

actually being due to this, when it is really
due to that other component of the composite.

But that seems like a very weak kind of unintended consequence.

DR. DE LONG: I would think the more relevant concern would be that there is demonstration that one of the composites is doing harm. What do you? As was mentioned, do you throw it out and still use the rest of the components or do you go back to the drawing board for the composite?

DR. ROMANO: The options here, it seems, are that we could, at one extreme we could say that 4c simply doesn't apply to composites because composites, in general, dilute, as Alan has stated very well, dilute problems, unintended consequences that come from a single component. So that would be extreme to say 4c just doesn't apply to composites.

But if we don't go to that extreme, then how do we rewrite or refrain 4c so that it seems applicable to composites?
DR. BURSTIN: I guess I am sort of lost. I mean to me it is what it is. I think it still applies. There may be some special circumstances in understanding what is causing it but to me, again if the composite submitted is required to do 4c, this is the general criteria. This is not unique to composites, but there may be issues where perhaps in the guidance the considerations you would say to the committee, if there is evidence of unintended consequences related to composite, the developer should provide information trying to locate the locus of the issue within the composite. Something like that. I don't know that we need to get into more detail than that.

DR. ROMANO: Right. So maybe then the way to frame this is because 4a and 4b, as I read it are about the composite as a whole, 4c really is going to be about the components of the composite. So maybe the argument in 4c is that if there is evidence of unintended
negative consequences for one of the
components of the composite, then the
developer should explain how that is handled
or justify why that component still belongs in
the composite. Is that --

DR. SHAHIAN: That's good.

DR. BURSTIN: Good work.

DR. ROMANO: Okay, so 4a and 4b
basically are applicable to the composite as
a whole; 4c is applied to individual
components with simply a rationale statement
or an argument the developer has to make. Is
that where we are? Okay.

And in some cases the argument
might be that the weighting is really low. In
other cases the argument might be that it is
balanced by something else within the
composite. In other cases it might be that
they dropped the measure, in the case of
pneumonia, I think, where they increased the
time limit actually. That is one that they
dealt with it.
Okay, so is that it for usability in use? All right.

So we go on to -- then end is in sight. So feasibility. We are back up here with the list. Are the feasibility criteria changing?

DR. BURSTIN: Not yet.

DR. ROMANO: Not yet. There may be a process in the future for changing feasibility criteria?

DR. BURSTIN: We are actually in the middle of a project right now looking at eMeasure feasibility and how to assess it. So I think it is going to potentially have some spillover into -- at some point we are going to have to re-do feasibility. It is such a new and different world than it was when we wrote it five years ago.

DR. ROMANO: I mean to me the key part of it is really the first part. On the left column it is the extent to which the required data are readily available and
retrievable without undue burden and can be
implemented for performance measurement.
Everything under there is just kind of
operationalizing that general principle.

So let's look at 4a, b, d, and e
on the right side and see if anybody has any
comments or suggestions related to those
criteria as they are applied to composite
measures.

MS. PAGET: Patrick, I have a
comment and this might be more a general NQF
comment. But under 4a where we state
routinely generated, right now it is not
routine that we generate patient-reported data
and maybe, Helen, the PRO project is going to
inform these criteria ultimately. But it
seems to me that again turning back to the
vision that these composite measures are going
to be used in areas like purchasing and so
forth that we are still making it fairly easy.
It is existing data sources that you can
readily get to. And I just didn't know
whether it is an opportunity for us to put in
some different language that recognizes that
we hope to get further along with patient-
reported data.

DR. BURSTIN: That is definitely
something you will see woven all through the
PRO report. I'm just not sure I know where to
put it here but I hear what you are saying.

DR. ROMANO: But it is a crucial
point because this feasibility framework
doesn't incorporate patient-reported data at
all.

I mean I think the original
intent, for example, for the clinical measures
was that we didn't want -- NQF doesn't want
measures that actually force clinicians to do
tests that they are not otherwise doing. You
know to send patients to x-ray or do lab
tests.

So they should be things that are
routinely generated concurrent with and as a
byproduct of care processes because we are not
trying to influence the care process by
forcing people to measure things that they
wouldn't otherwise measure.

DR. KAPLAN: How do you get that
from the language that is there?

DR. ROMANO: Well there could be a
footnote, maybe to clarify what is meant by
exclusion there is not forcing people to do
tests or to spend money that they wouldn't
otherwise do.

DR. BURSTIN: I mean it is really
about the burden of measurement is really what
it is saying. Is the juice worth the squeeze.

DR. ROMANO: The burden, including
the harm of measurement. I mean potentially
you could create a situation where you are
actually causing harm by having people do
things that aren't a routine part of the care.

DR. BURSTIN: I've never really
thought about it that way, Patrick. I have
always thought about it more as just as the
burden of measurement.
DR. CHASE: I would just say the exception to that may be where the collection of data is actually part of the care process and collection of PHQ-9 for patients with depression is argued to be one of those. I don't know if that is what you are --

DR. ROMANO: Well the counter argument for example is a lot of people have used albumin measures in hospitalized patients because it is a strong predictor of various adverse outcomes. But the argument is that testing albumin leads people to do strange things like giving albumin to try to raise the albumin value, which is not therapeutic.

DR. BURSTIN: The only thing again is you go back to the report last time around since it was literally one paragraph on feasibility just indicate that since composites are more complex, the data collection methods are going to need to make sure you can pull in all the different components.
So the data collection strategy for obtaining all required components needs to be combined in the composite measure to demonstrate it is feasible. So if you are getting data from all different data sources, all that should come into play as you are looking at the feasibility of composite. Other than that, there was nothing distinct and different.

DR. ZASLAVSKY: Are there any issues here about feasibility that are emergent with composite measures in particular? All I can think of is whether maybe you would want to be able to get all of the measures at the same time or something like that, whether there would be -- I don't know whether it is even worth mentioning.

DR. ROMANO: I mean it would seem that in general there is presumption that in order for the composite to be feasible that all of the individual components are feasible. Right?
DR. BURSTIN: Yes.

DR. ROMANO: Because the feasibility will be driven by the least feasible component.

DR. ZASLAVSKY: What if you have had a composite that required you to link at the patient level things which come from different sources? That is about the only emergent feasibility issue that I can think of.

DR. ROMANO: It's a good question. So how would you frame that in this context, that developers should identify any feasibility issues related to the composite above and beyond the individual measures?

DR. ZASLAVSKY: I guess if we just repeat the same language for the composite that we have for the individual measures, that would cover it.

DR. SHAHIAN: The same deficiency does apply though to apply -- the same problem does potentially apply to individual measures.
We just did a 30-day CABG readmission measure where we linked the STS clinical database with MedPAR. And that of course was done long after the delivery of care. So it is a general problem, it is not just a composite problem.

DR. ROMANO: So I'm not hearing any specific suggestions about things to change in 4a, b, d, and e here recognizing that probably some of the ongoing work related to patient-reported outcome measures and related to eMeasures should drive a more general rethinking of this entire category.

DR. GOLDSTEIN: I was going to say, and this isn't I think specific for composites but relates to the individuals, something about auditing the data but if they have like a plan how they are going to audit the data or does it need to be audited or does it need to be checked against other measures. For some of our systems, in particular for high stake systems we spend -- that is a huge,
huge area that we spend a lot of time focusing
is it audited. Is it audited by independent
review? And if it is not, are there other
sources of information to validate the data,
especially if it is used for pay-for-
performance, eventually any of these measures
that incentives for gaming increase.

So we put penalties if they do
anything to bias the rates there is a huge
penalty in terms of what we publically report.

DR. BURSTIN: That is interesting.

I don't think of audibility or auditing under
feasibility, more so under validity. And
actually it has been something that is, you
know, it's too bad David has left the room,
that is a big part of what the STS measures
have brought forward is the percent sample and
auditing et cetera.

DR. ROMANO: I have been having
trouble with 4d here in general terms. It is
a hard sentence to parse, to borrow Alan's
description earlier. It is just difficult
because the data on inaccuracies and errors and unintended consequences are being described under other evaluation criteria. So this is somehow about the ability to do it. It just seems odd in this context.

DR. KAPLAN: I'm confused about b, d, and e buy you for the composite that you don't already have for the individual items. I mean they look to me identical except -- yes, so why do you need them?

DR. BURSTIN: Because the idea would be sometimes you only just have a composite and are they applicable or not. So all this is saying is yes, those are applicable, same wording applies.

DR. KAPLAN: Okay so b then needs all the other junk that is in 4b on the other side? Because right now it is written as assertional versus aspirational. Right? Because some of the data like the patient experience data, supposing you were
creating a composite that included patient
reported data with other kind of data, that is
not in the electronic form right now.

DR. BURSTIN: You are right. I
don't understand why the difference.

DR. ROMANO: So we have raised a
number of questions about these criteria in
general. Any final comments, suggestions
regarding the feasibility criteria?

Okay, great! So I think we want
to have another opportunity for public
comment. We, I think, have basically
discarded the idea of a taxonomy, per se.

DR. BURSTIN: Yes.

DR. ROMANO: So we are not going
to discuss that.

Do we want to talk at all -- in
terms of the submission form, we decided sort
of not to go through the submission form in
detail because that would be a very tedious
exercise. It doesn't add anything.

But we do want to give NQF staff
guidance about how to redesign or tailor the
submission form.

And I guess what I would like to hear a little bit of discussion about maybe
for five minutes is do we -- I think earlier
we decided that there isn't a bright line
between composites and other types of measures
and that we weren't comfortable with having an
entirely different process for composites than
for other types of measures. In principle
many of the measures that NQF has endorsed as
individual measures could be viewed as
composites and some of the measures that have
been endorsed as composites could be viewed as
individual measures, depending on your
perspective. So what matters a lot more is
the measurement concept or the constructural
framework.

So how does this relate to the
form? Are we going to ask people to declare
their measure as a composite or not? Are we
going to have questions embedded within the
forum that says that if you are building your
measure on other measures then you must go
over and answer this additional set of
questions? Any thoughts about how to
structure that in general so that the review
process is uniform across steering committees?

DR. ZASLAVSKY: I guess one thing
would be to have an option to indicate which
existing measures are incorporated into your
measure. So you sort of incorporate by
reference the whatever approval has been done.

I guess the other side of it is
what -- the thing we have been struggling with
is when do you kick in these additional
criteria? In a sense, anytime you rely on
existing measures, then the new criteria for
combining things should kick in. The question
is, are there situations where you are not
relying on any existing approved measures but
you would still want to bring in these
criteria. That is where we run into this
territory where we have a hard time drawing
the line between what is the composite measure and what is just a single measure that happens to incorporate a bunch of different elements?

So you know I think we have heard the argument there from several people, including myself that a lot of these criteria we have introduced for the composite measures actually should apply to things that are called single measures but that combine a number of different things. So I am not sure how we address this questions of whether to just make this part of anything that isn't really just completely one thing or do we really try to distinguish the subset for which this is applicable?

DR. ROMANO: Very well described. Anyone able to answer Alan's question?

DR. DUNTON: Well if you don't have an opportunity to declare it as a composite, then it doesn't trigger in some of the other things which may or may not be added, such as a purpose in the conceptual
model. And so I'm not sure you can get away forever with not defining what a composite is, although I don't have a solution.

DR. KAPLAN: I have a separate issue, which is the issue of harmonization with related measures and then the competing measures things, e2 and e3. Harmonization is kind of the activity that is involved in creating the composite measures. You are pulling things together that theoretically assess some other construct. So you will try --

DR. BURSTIN: No, it is different. So what we are referring to there is essentially comparing what you have inside your composite to other measures.

DR. KAPLAN: Other parallel composites.

DR. BURSTIN: Other similar constructs.

DR. KAPLAN: How often is that going to happen?
DR. BURSTIN: Not necessarily just constructs, the individual measures happens half the time in our lives.

DR. KAPLAN: Well I know it does with individual measures. I am trying to get at so say you have -- forgive me for math again, but you have your four test questions in algebra. Now somebody else has four other questions that they assert measure algebra. They are very different questions. They both measure algebra. So you are asserting that that other test that has these components is different from my measure in the following ways, blah, blah, blah, blah, blah? Or are you suggesting that now I make an eight-item version of my algebra measure? I am confused with respect to the composite issue. I am a little confused about how this plays out.

DR. ROMANO: I think it ties with the broader context and the concern about measurement burden.

So just to cite one example from
AHRQ's composite. So there is a patient safety composite that includes a measure of central venous catheter associated infections. It is based on administrative data. At the time that was chosen because it was a feasible measure to include in that composite. But there is a practical problem which is now we have another measure which is in Hospital Compare which is based on National Healthcare Safety Network reporting, using specific definitions that come from the CDC. So that is not harmonized.

So we have a measure of CLABSI within the PSI composite that is not harmonized with a different measure of CLABSI that is also publicly reported at a national level that is defined by CDC.

So one solution to that approach, for example, would be for AHRQ to reconstruct its composite to actually include the CDC measure instead of the measure based on administrative data.
Another approach would be to justify why they can't be harmonized based on some superiority of the administrative database measure for this particular application. But I think that is what we are getting at. Does that make sense?

DR. KAPLAN: Yes, it just has a different flavor when you are talking about a composite versus -- I get it at the individual measures level. You guys are trying to make sure that you are both coming at this from the same direction. You end up agreeing on denominators and all that other stuff.

What is a little different is now you have got three measures that are the same and one measure now that looks oddball and you are going to figure out how to grab that in. What is it going to do to the composite which is a little bit different issue when you are talking about harmonization.

DR. CHASE: So the only thing I would hope about this is that we don't put a
different burden on composite measures just because we can. Because harmonization should affect, I think, all measures as they come through and I think that is the intent. So when other measures come renewed as well, so in both cases it wouldn't just put it on to the composite measure developers to align with what others have already done.

      DR. ROMANO: I am not sure that this issue has arisen but theoretically it could arise that different developers could come in with different measures of what they claim is the same thing.

      DR. DE LONG: Diabetes. We have already done it in diabetes.

      DR. ROMANO: Right. So then -- but actually NQF had decided in that case to endorse both, presumably based on some rich discussion and some rationale.

      DR. BIRKMEYER: But also --

      DR. BURSTIN: Rich! I think it was rich.
DR. BIRKMEYER: But also just given the possibility that sort of NQF views its role as deciding when a measure is rigorous and even if there are two things that are ostensibly measuring the same things with slightly different tradeoffs, there is a marketplace public and private that will sort out which one ultimately gets used.

DR. BURSTIN: Although in that case, the individual components were harmonized. So where the measures overlapped the science was harmonized for the level of blood pressure control, LDL, and A1c.

DR. ROMANO: Anyway, there is a separate process I think that AHRQ is currently -- that NQF is currently undertaking internally to review its approaches to harmonization. So that is on the back burner for today.

DR. BURSTIN: I'm not getting a whole sense of energy left to do the composite.
DR. ROMANO: No, I think people are fading.

DR. BURSTIN: I think it might be more useful once we have tried to write up the criteria to come back. I do think that, Karen and I talked about this a little bit earlier that I guess Karen has also suggested that in some ways since there is so much uncertainty that likely the best approach would be to not have a whole separate composite submission form. Have this -- just like do if it is an outcome, is there risk adjustment. If it is a composite, you answer this series of questions. And maybe the question is going to be do we allow the developer to self-trigger is it a composite. We can have our guidance what we think they are. But maybe part of this is also an assessment on the part of the committee of well we actually think this is a composite, you need to go back and answer those questions.

So we will play with that a bit.
DR. ROMANO: And I think Alan's idea is definitely useful of bringing in other measures by reference and where that cannot be done, then the question is does the developer have to submit separate information about individual components that have not been separately endorsed? So would there be a requirement for a separate submission form for individual measures that are not currently endorsed?

DR. BURSTIN: I think some of that, again, comes back to the clarity of what we are actually saying has to be submitted for the components.

At least I continue to hear that we talked about the importance to measure report, still needing to be there for both the overall and the components and not always for the -- it is a big waste to have the form have sort of pop out boxes if it is -- you know, be able to fill in the additional information required.
But some of this will come back
down to the developer as well. Do you want to
put forward the new components within your
composite as stand-alones for endorsement?
And if that is the case, then something else
would pop up that they would complete. But
that may not always be the case.

DR. ROMANO: But if you don't,
then there should be a way to invoke a simpler
process.

DR. BURSTIN: Yes.

DR. ROMANO: Yes.

DR. ZASLAVSKY: So I don't think
there are many of the things that are in the
overall criteria that aren't in the component
criteria but there are a few that are at least
softened or modified so you maybe have a
modified version of the form.

DR. ROMANO: Should we open to
public comment?

DR. BURSTIN: Yes.

DR. ROMANO: Is there anyone else
in the room who would like to comment? In the meantime, we can open the telephone lines.

OPERATOR: At this time, if you would like to ask a question, please press *, then number 1. And there are no questions.

DR. ROMANO: Going, going, gone.

DR. BURSTIN: We're gone.

DR. ROMANO: So the public comment period is closed for right now. Next steps.

MS. MUNTHALI: So in terms of our next steps, the panel will reconvene via conference call in about two weeks actually, November 15th and we will follow up on any issues that we have.

In the process, our team is going to start drafting the technical report with your recommendations. You may be seeing some emails for clarification. We have been taking notes. We also have Eric in the back who is our court reporter and so we will be looking through all of that. I am going through the recording as well.
Also after we draft a report, we will give you an opportunity to look at the report before we post it on the NQF website for our member and public comment and that is a 30-day period.

But because this is very important work that will influence evaluation guidance for composites, we want to have our Consensus Standards Approval Committee review the draft report before we receive comments from our members and public. We want them to start looking at the material, get comfortable with the issues that you have raised. And so we will go to the CSAC on December 10th and we will have a call to adjudicate the comments that we received during the comment period on January third and go back to CSAC with final recommendations, which also considers our member and public comment on January 8th. And we hope to have Board ratification a few weeks after that, hopefully by the end of January.

I don't know if you have any
questions on our time line. And of course you can contact Karen Johnson, Karen Pace, or myself. We are on SharePoint. I hope you guys are using it. I'm sorry for any issues you have had if you had any issues.

And we just wanted to thank Liz and Patrick for leading such a great discussion today and to thank all of you for coming and we look forward to communicating again online and via phone and safe travel back home.

DR. ROMANO: Just a question just to clarify. Any changes to the submission form and procedures would follow this entire process?

DR. BURSTIN: Yes.

DR. ROMANO: Or would they be done concurrently? We would follow the process?

MS. MUNTHALI: Yes.

DR. ZASLAVSKY: Will you be able to tell us soon what time the next call is at because some of our schedules are filling up.
MS. MUNTHALI: Yes, we are sending
that to you. I can tell you it will be in the
afternoon but I don't know the exact time. It
will be a two-hour block. We have been trying
to make sure we accommodate time for the --

DR. BURSTIN: Three to five.

MS. MUNTHALI: -- three to five, right?

DR. BURSTIN: Yes, 3:00 to 5:00 on

November 15th.

MS. MUNTHALI: So we will send that
in an email and post it on SharePoint as well.

DR. ROMANO: Okay and if there are
any other suggestions for NQF or for me and
Liz, please let us know. Thanks again.

DR. BURSTIN: Thank you for such
able facilitation of a tough topic. And thank
you all. This was an interesting
methodologically challenging day. Thank you
for your brain power for the day.

(Whereupon, the above-entitled

matter went off the record at 4:13 p.m.)
<table>
<thead>
<tr>
<th>ability</th>
<th>97:21</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolutely</td>
<td>120:12</td>
</tr>
<tr>
<td>24:11</td>
<td>135:19</td>
</tr>
<tr>
<td>178:19</td>
<td>186:19</td>
</tr>
<tr>
<td>236:4</td>
<td>300:14</td>
</tr>
<tr>
<td>332:12</td>
<td>361:5</td>
</tr>
<tr>
<td>ABIM</td>
<td>270:1</td>
</tr>
<tr>
<td>able</td>
<td>29:11</td>
</tr>
<tr>
<td>12:31:5</td>
<td>45:14</td>
</tr>
<tr>
<td>47:18</td>
<td>52:1</td>
</tr>
<tr>
<td>63:21</td>
<td>70:6</td>
</tr>
<tr>
<td>78:7</td>
<td>82:5</td>
</tr>
<tr>
<td>121:5</td>
<td>124:17</td>
</tr>
<tr>
<td>144:17</td>
<td>179:8</td>
</tr>
<tr>
<td>183:2</td>
<td>10</td>
</tr>
<tr>
<td>185:2</td>
<td>188:8</td>
</tr>
<tr>
<td>191:10</td>
<td>197:21</td>
</tr>
<tr>
<td>202:9</td>
<td>211:9</td>
</tr>
<tr>
<td>248:7</td>
<td>252:13</td>
</tr>
<tr>
<td>262:13</td>
<td>277:22</td>
</tr>
<tr>
<td>279:9</td>
<td>285:3</td>
</tr>
<tr>
<td>329:20</td>
<td>330:10:19</td>
</tr>
<tr>
<td>331:4</td>
<td>337:22</td>
</tr>
<tr>
<td>338:2</td>
<td>13</td>
</tr>
<tr>
<td>357:14</td>
<td>365:17</td>
</tr>
<tr>
<td>373:21</td>
<td>377:20</td>
</tr>
<tr>
<td>378:17</td>
<td></td>
</tr>
<tr>
<td>above-entitled</td>
<td>152:1</td>
</tr>
<tr>
<td>219:4</td>
<td></td>
</tr>
<tr>
<td>324:15</td>
<td>378:21</td>
</tr>
<tr>
<td>absence</td>
<td>19:3</td>
</tr>
<tr>
<td>absolutely</td>
<td>24:11</td>
</tr>
<tr>
<td>72:1</td>
<td>84:2</td>
</tr>
<tr>
<td>119:4</td>
<td>240:12</td>
</tr>
<tr>
<td>245:10</td>
<td>251:5</td>
</tr>
<tr>
<td>abstract</td>
<td>289:18</td>
</tr>
<tr>
<td>ACC</td>
<td>149:15</td>
</tr>
<tr>
<td>accept</td>
<td>122:7</td>
</tr>
<tr>
<td>191:18</td>
<td></td>
</tr>
<tr>
<td>acceptability</td>
<td>13:3</td>
</tr>
<tr>
<td>15:18</td>
<td></td>
</tr>
<tr>
<td>18:3</td>
<td></td>
</tr>
<tr>
<td>58:20</td>
<td>94:2</td>
</tr>
<tr>
<td>217:4</td>
<td>239:11</td>
</tr>
<tr>
<td>265:11</td>
<td>12</td>
</tr>
<tr>
<td>acceptable</td>
<td>85:12</td>
</tr>
<tr>
<td>85:12</td>
<td>108:14</td>
</tr>
<tr>
<td>111:21</td>
<td>177:2</td>
</tr>
<tr>
<td>213:1</td>
<td>233:19</td>
</tr>
</tbody>
</table>

283:8,9 284:3
323:16
accepted 122:12
accepting 109:18
access 104:12
acceptability 196:6
accommodate 378:5
accomplish 18:12
25:20 42:19
238:16
accomplishing 100:12
accountability 16:5
20:2,4,17 22:5,16
22:19,20 135:14
328:22 335:15
341:1,19
accredit 99:5
accumulating 346:20
accuracy 90:6
272:10 290:3,5
accurate 101:14
accuse 76:13
achieve 79:4
340:20 341:8
achieved 334:12
achieving 336:5
acknowledge 309:15 313:20
ACO 49:10
acronym 199:7
210:14
act 72:8
acting 8:7 306:7
action 122:17
activation 346:14
activities 15:18
334:7
activity 309:21
310:13 366:8
actual 18:20 19:17
37:11 144:17
335:20 343:15
acute 235:16,20
ad 114:18
adapted 229:14
265:17 335:13
add 77:2 80:7
102:11 112:17,22
113:3 133:6,6,8
133:12 148:1
183:14 190:10
201:3 207:3
209:21 215:15
248:16,21 249:1
254:15 310:7
311:12,12 312:18
313:2 319:1
326:11 362:21
added 37:15
152:18 190:22
214:2 221:1
228:20 229:6
241:16 244:9
265:21 365:22
adding 112:5,19
113:14 118:8
120:16 148:17
188:17 190:19
231:8 251:16
256:22 291:7
301:18 309:21
addition 298:9
additional 35:7
66:10 111:19
164:4 168:12
175:6 204:13
205:1 222:6 229:5
241:12,22 255:14
266:15 326:2
364:3,14 373:21
Additionally 334:10
additive 213:21
250:19
address 73:6 136:1
146:2 365:11
addresses 136:10
adds 80:21 207:6,8
207:11 258:5
321:14
adenoma 61:6
adequate 134:12
adequately 51:8
Adjourn 3:22
adjudicate 376:15
adjusted 58:19
60:18 177:5
adjuster 98:11
adjustment 55:16
172:20 173:4
273:13 372:12
administering 115:10
administration 2:8
15:7 346:11
administrative 368:4,22 369:3
administratively 167:2
admission 344:5
admissions 209:6
advanced 146:8
advantage 10:22
11:3 89:12,13
104:18
advantageous 124:6 197:9
advantages 345:9
adverse 58:16
356:11
advice 92:7
advise 11:11
advised 14:5
advocate 14:6
185:15 206:11
248:1
advocates 314:9
Affairs 8:6
affect 245:22
274:12 319:7,8
370:3
affirmative 343:7
afternoon 111:10
175:1 212:19
220:6 324:12
326:12 378:3
age 177:5
agencies 216:17
agenda 19:5,9
222:1
agnostic 183:7
agnostically 90:17
ago 70:17 227:3
228:18 244:22
306:12 352:18
agree 47:2 78:16
79:8 81:13 84:11
90:8 93:8 94:5
96:10 97:11
117:19 118:2
119:4 123:22
138:22 141:12
146:15 163:9
199:14 227:20
247:19 263:5
282:11 294:2
299:14 312:12
319:17 320:14
338:4
agreed 55:18 224:6
224:15 264:6
303:5 314:17
315:1 339:8
agreeing 316:10
369:12
agreement 125:9
223:7
ahead 16:16 65:7
73:2 152:4 153:3
196:4
AHQR 6:14 13:19
13:22 14:13 17:18
46:7 58:3 72:12
87:8 269:3 318:3
331:11 368:19
371:15
AHQR's 368:1
aim 100:9
aims 100:12
airplanes 112:6
120:16
Alan 2:9 8:14 27:8
82:10 84:5 97:9
102:22 130:11
154:12 170:1

Neal R. Gross & Co., Inc.
202-234-4433
| 1a 228:19,21 | 253:17 263:21 | 264:6,7,10 |
| 1b 228:22 253:22 | 255:8,12 262:22 | 264:12 |
| 1c 229:2 253:16 | 254:5,8,13 258:20 | 260:5 264:11 | 275:1,12 |
| 1d 228:20 229:6,12 | 229:18,19 230:7 | 233:2 253:3 |
| 1e 229:9,13,19 | 230:3 |
| 1:10 219:6 220:2 |
| 10th 376:14 |
| 10:55 152:2 |
| 100 144:18 340:20 | 341:2,8 |
| 1030 1:12 |
| 11 272:21 273:1,8 |
| 11:13 152:3 |
| 12 108:5 134:16 | 136:16 138:9 | 325:8,9,15 |
| 12-page 221:5,8 |
| 12:20 219:5 |
| 140/80 193:16 | 256:14 |
| 140/90 193:17 | 260:11 |
| 15 50:6 345:2 |
| 15th 1:12 375:13 | 378:10 |
| 16 3:9 30:20 |
| 19 191:3,8,10,15,18 |
| 1900 10:7 |
| 2 1:9 100:4 265:10 | 273:4 296:9 |
| 2a 265:13,16,20,21 |
| 2b 266:3,3 272:19 |
| 2b.1 272:8,8 273:9 |
| 2b.2 272:8,9 273:10 |
| 2b.3 273:8,11 |
| 2b.4 272:14 273:12 |
| 2b.5 273:13 |
| 2b.6 273:15 |
| 2c 272:2 |
| 2f 291:21 298:20 |
| 2h 292:10 |
| 3 273:4 |
| 3d 326:2,4 |
| 3e 326:2 |
| 3:00 378:9 |
| 3:01 324:17 |
| 3:15 324:14 |
| 3:16 324:18 |
| 30 202:1 216:20 |
| 30-day 38:17,20 |
| 41:5,11 359:1 |
| 359:5 |
| 36 129:10,11 |
| 362 3:19 |
| 374 3:20 |
| 375 3:21 |
| 378 3:22 |
| 3d 326:2 |
| 3e 326:2 |
| 3f 326:2,4 |
| 3g 326:2 |
| 3h 326:2,4 |
| 3i 326:2,4 |
| 3j 326:2,4 |
| 3k 326:2,4 |
| 3l 326:2,4 |
| 4 3:3 222:2,5,15 | 272:15 273:4 |
| 325:16 |
| 4a 335:9,14,19 |
| 339:8 350:18 |
| 351:8 353:5,12 |
| 359:9 |
| 4b 336:2,3,18,19 |
| 339:9,10 350:18 |
| 351:8 361:18 |
| 4c 336:17 339:15 |
| 349:13,18,21 |
| 350:6,20,21 |
| 351:10 |
| 4d 360:20 |
| 4:13 378:22 |
| 400 28:13,14 |
| 5 3:4,4,7 273:4 |
| 5:00 378:9 |
| 50 104:19 111:1,2,5 |
| 111:17 118:16 |
| 248:22 301:18 |
| 50-something |
| 113:19 |
CERTIFICATE

This is to certify that the foregoing transcript

In the matter of: Composite Measure Evaluation
Guidance Project Expert Panel

Before: NQF

Date: 11-02-12

Place: Washington, DC

was duly recorded and accurately transcribed under
my direction; further, that said transcript is a
true and accurate record of the proceedings.

-----------------------------------
Court Reporter