

## NATIONAL QUALITY FORUM

**Moderator: (Karen Pace)**  
**January 3, 2013**  
**1:00 p.m. ET**

Operator: Welcome to the conference. Please note today's call is being recorded.  
Please standby.

(Karen Pace): (Nathalie)?

Operator: Yes, ma'am?

(Karen Pace): Hi, is that our queue to start or is there...

Operator: Yes, ma'am.

(Karen Pace): ... OK. Thank you so much. OK, this is (Karen Pace), and as I said, (Karen Johnson) and (Elisa Munthali) are here with me at NQF. I'm just going to quickly, once again, say who's on the call. I think almost everyone is here. We have our co-chairs, (Patrick Romano) and (Liz DeLong). We also have (John Birkmeyer), (Dale Bratzler), (Nancy Dunton), (Liz Goldstein), (Sherrie Kaplan), (Dave Shahian), (Steve Wright) and (Alan Zavslavsky).

So, I think we'll go ahead and begin. We have quite a bit to discuss. I want to thank you all for your – the TEP for all of your reviews and suggested changes. We've tried to incorporate all of those changes to the best of our ability but certainly if we've gotten something wrong or something that needs further discussion we are more than welcome to hear that from you and encourage you to definitely let us know.

The purpose of the call today is to – (two fold) really to review the public comments we received and also to, you know, make sure that we had made

the appropriate changes based on the TEP additional review of the posted document as well.

And we'd like to really focus today's call on making sure that we have the key recommendations and criteria correct because those are ultimately the things that will be used in the future to help us evaluate composite performance measures then – and certainly, you know, additional edit to the report. We welcome but we can maybe do some of those offline so that we make sure that we use this time to discuss and resolve any issue.

Before I get started, I want to see if (Patrick) or (Liz) wants to make any overarching comments or suggestions for the call.

(Patrick Romano): Well, I don't actually. I mean, I have to be honest that with the holidays and so forth, I really haven't a chance to look at what you folks sent out yesterday. So, I apologize for that in advance. But we'll look forward to your orienting us and I think that, you know, we may have some additional comments as we talk about things and after this call as well.

(Karen Pace): OK, great! (Liz)?

(Elizabeth DeLong): Likewise, I haven't been able to get through all the work you've done, but you've (teed) things up nicely. So, I think it'll be a very productive discussion.

(Karen Pace): OK. So, thank you all, and I'll just ask if any of the other TEP members have any kind of key burning issues that we need to make sure that we've got on the – on the agenda to get to today.

OK, why don't we proceed then, and we will certainly feel free to interject at any point, but we try to, in the briefing memo, we identified kind of five-key themes for discussion, and we'll go through those first and then certainly if there are any additional comments or address to responses that you'd like address or any other issues in the report, we'll be glad to do that. But one of the first things we wanted to discuss is the definition of a composite and clearly identifying what types of measure should be classified as a composite performance measure.

And on page four of the report, we have the definitions and actually this was an area we had several suggestions from the TEP to rewrite that paragraph. So, we need to confirm that we have the correct definition of a composite performance measure and if you're satisfied with the rewrite of that paragraph about a composite performance measure.

So, basically we still have that a composite performance measure is a combination of two or more individual performance measures resulting in a single measure that result with a single score. So – and then we have the rewrite of that paragraph. And again if we find that we need more time, we will certainly adjust as we go through the call today. If possible, our timeline was to take final recommendations to the CSAC on next Tuesday, but I think that will depend on whether we identify things that need further review and discussion by the TEP as we go through today.

Male: (Karen)?

(Karen Pace): Yes.

(Patrick Romano): Just a general question, I'm just curious why the timeline on this project seems so incredibly aggressive. You know, given the holidays, both Thanksgiving and the Christmas and New Year holidays, it seems like the timeline is sort of marching forward without respecting those natural time breaks. I was just curious if there's some particular pressure that this has to get done next week or...

(Karen Pace): It's a good question. From the beginning, this project was – it has to do with our funding cycle, and there was a lot of pressure to get this done by the end of the year which obviously wasn't possible, so then it was moved to the first CSAC meeting of the year but, you know, we definitely hear you, and we had some discussions with Heidi this morning and that's why I mentioned that, you know, and maybe we should have that discussion first.

If the TEP really feels that you haven't had adequate time to look at any of these latest materials, we can certainly, you know, have gotten approval that we could push this to the February CSAC meeting which would give us a little

more breathing room. So, maybe I should just first hear whether, you know, everyone is feeling that pressure and if that we should kind of move to plan B or how others feel about that.

(Elizabeth DeLong): Actually, I think, Pat – this is (Liz). I think (Patrick) makes a good point but without going through the document to elicit comments, I'm not sure we have a good sense of whether we should delay. I mean, I think we all feel probably we will but maybe your agenda was put in appropriate chronology.

(Alan Zavslavsky): This is (Alan). I had some more reaction. I think we should go on with our business and, you know, we're not – just keep in mind that we're not necessarily going over this any kind of quiz. If we can't resolve every issue but do our best to do it and then decide at the end whether we still have enough unresolved issues that we really need more time.

(Patrick Romano): That sounds fair to me.

(Karen Pace): OK. Thank you all. So, we'll – and Elisa heads up on the Webinar the section of the report that we're talking about in terms of the definition in rewrite.

So, the next – so just pause there before I move on and then we can, you know, definitely we can always come back if some of these things that'd be interrelated.

(Alan Zavslavsky): Can I ask about that definition? I (know) this (inaudible) complicated, you know, at (site) line there between using anything that puts together more than thing and quite of a composite and requiring, you know, with the other extreme that every composite made of things that could stand by themselves, but is this phase each of us can individually be used to assess quality of health care services. Is that the lower standard than being potentially endorsable?

I guess someone ask...

(Karen Pace): Yes, I guess – I'm glad you pointed out. That was one of our questions. Someone had suggested that we add that language, each of which can individually be used to assess the quality of health care services and wanted to

make sure that it didn't kind of lead to the conclusion that each had to be individually endorsed because your recommendations have been that we don't require that each component be individually endorsed.

So that is one of our questions if whether we need that phrase in there, and really wanted to have some discussion about that.

(Alan Zavslavsky): So, I'm trying to think of an example. You know, we have for example the (CAPs) composite which is near and dear to me which I think he could say that they are, you know, each item in the composite could be considered a measure of health care quality (inaudible) (upward) assessments of their quality.

(Karen Pace): Right and the (CAPs) is also – the (CAPs) composite instrument is a composite but the performance measures relate to one domain. So, this is one of those areas where it gets a little confusing in what we've tried to identify is that there's a distinction between a composite performance measure and a composite instrument or scale. You know, the terminology of composite is used in a variety of ways. But I think what we're trying to get at here is about the performance measure and a composite scale or instrument may or may not end up being used in a composite performance measure depending on how, you know, it's actually used.

(Alan Zavslavsky): Let's take another example. One of these all or nothing measures.

(Karen Pace): Right.

(Alan Zavslavsky): So, you know, administering antibiotics, checking for allergies, you know, and those eight different things you might do when you're doing surgery. Each of them is a measure of performance (from those) there any more or less of a composite then these four items and they're getting clear quickly composited (test).

Male: Well, I would say, I like this wording because ultimately it's up to the measure developer to decide what constitutes an individual performance measure. So, if you were to say that those component processes prescribing the correct antibiotic, prescribing it on time, (stopping) it on time so forth. If

you were to say that each of those is an individual measure, then the definition is satisfied similarly with (CAPs). If you were to say that each of those components is a measure that measures something of interest, it could stand on it's own as a measure then (satisfy) the definition.

On the other hand, if you say that these four questions are really just ways of getting at a single concept and it doesn't make sense to look at the four questions individually because they're all designed to get at the same concept. That would be different. That would be – that wouldn't satisfy the definition.

What is...

(Alan Zavslavsky): I'm just not sure of that (inaudible)...

Female: Yes.

(Alan Zavslavsky): ... to the year that committee or, you know, the (staff) taking an application or the persons to meeting the measure.

(Sherrie Kaplan): Yes, I am confused because for me the language performance means the act of doing something, and you don't think of, you know, patient health status as a performance of the health care system. It certainly could be used in that context, but if that what's meant by performance, I think that's the thing I'm getting hang up on because yes while all of that's (inaudible) that individual like these three or four items that measure communication in (CAPs) are intended to measure communication.

Individually, the point as this pointed out on page 10 of doing composite measure is to improve precision around the construct, and if you're trying to measure a single construct with multiple items. If that what's – is this paragraph really helping to delineate the difference between that and making a composite out of ordering annual hemoglobin A1c, is checking feet annually doing eyes – blah-blah-blah. All of which are intended to reflect quality of diabetes care.

So, I don't – this doesn't help me. Maybe this is using language that I'm not (inaudible) and maybe this is (Alvan Feinstein) again messing out with us

with clinimetrics versus psychometrics but this language (is) serious in helping me.

(Karen Pace): (Sherrie) are you talking about the – which language are you talking about now that – on page four about the insert of each of which can individually be used to assess the quality?

(Sherrie Kaplan): Yes, I get that because if to me that – like if you should do annual eye exams on diabetics (period), you know, that...

(Karen Pace): OK.

(Sherrie Kaplan): ... should be done. If you're creating a composite around diabetes care than adding more things to that composite help you understand whether or not the construct quality of diabetes care is being enhanced or not by the addition of more of those individual items. But if you're reflecting an individual construct using – I mean, for me, there's no difference in that.

The performance thing is what's hanging me up. I don't get the difference between multi-item measures of dimensions within (CAPs) versus multi-item measures of diabetes care.

Male: Well, I think perhaps – I mean – I mean, we talk about performance measures as measures of the performance of the health care system and that is generally use terminology in NQF documents. But maybe that parenthetical reference to (CAPs) and PHQ-9 is confusing because as – I think (Alan) and (Sherrie) have pointed out (CAPs) could be viewed as a composite performance measure.

So that would...

(Karen Pace): OK.

Male: ... fair if we took out that parenthetical statement and basically just, you know, just stuck with the theoretical concepts.

(Sherrie Kaplan): I think you should take it out...

(Karen Pace): OK.

(Sherrie Kaplan): ... cross out the parenthetical there.

(Karen Pace): OK. All right. So, we can I think do that easily and (leave) in the language we've just been talking about let's move on to a...

Male: (Karen)?

(Karen Pace): Yes.

Male: Supposing I might say is – respecting what (Sherrie) and Alan are saying, is that I think that multi-item instruments and scales about experience with care can in fact constitute performance measures. So...

(Karen Pace): Yes.

Male: ... somehow we need to tweak the first part of that paragraph, again, maybe simplifying it to make it clearer that well these things could be performance measures. I mean ultimately it's up to the developer to declare whether it's intended to be used as a performance measure or not.

(Karen Pace): But – yes, but I – and I guess this is where I need to bring in some of the language from the recent PRO project is that the instrument itself is not a performance measure. There has to be some aggregation of those individual level items into an actual performance measure. So, we've been trying to make clear that NQF doesn't endorse the (CAPs) or endorse the PHQ-9 what we endorse is performance measures that use data from those instruments.

So, the instrument itself does not equate to a performance measure unless you figured out how to aggregate that into a performance measure.

Male: Well, that's, of course, a very gray line because the performance measure is based on a particular instrument but anyway, respecting that line...

(Karen Pace): Yes.

Male: ... certainly, we can have the terminology consistent with what's in the PRO report.

(Karen Pace): Right. OK, good point. We will do that. Yes, it is a gray line because those measure specifications will indicate which instrument the data are coming from that's kind of analogous to the, for example, if you were using data from the Nursing Home (MDS) is a Home Health Oasis that it specifies, you know, the standardized data collection but we will kind of reference or use that terminology from the PRO report.

OK, the next thing that we wanted to – related to this or continuing from this is that the types of composite measure discussion. You know, we've all acknowledge then had in this report that, you know, there's no one way or good way to categorize the various approaches to composite measure development. However, we did receive some questions and comments and given that we have specific criteria that are going to be applied to composite performance measures.

The question has come up of, you know, how are we going to consistently ensure that those things that really are composite performance measures are identified as such so that those additional criteria are applied. So, it's not as simple as just letting each measure developer decide because that may not end up in consistent application of these additional criteria.

So, we wanted to, you know, bring this back to you. We drafted some language here which is new. So, obviously, it's something that's going to need your review and thoughts about but, you know, the question is, you know, how do we decide that the measure really should have the, you know, the additional analysis and evaluation that we're putting forth for composite performance measures.

So, what we tried here is to give some examples of what would be a composite – or discuss what is a composite performance measures would not – what would not be classified and then, you know, where we really had some questions that we thought merited just confirmation one way or the other from all of you.

One of the commenter's specifically questioned measures that use the shrinkage estimators or, you know, based on the hierarchical models that combine, you know, the provider performance plus an average performance whether it's the overall average or the average based on some other categorization.

The any or none kind of multiple complications that's another one that, you know, we've talked about, but we just need your thoughts about guidance from both – for both staff and steering committee members because I think, you know, that is what and that's also one of the issues that we identified in terms of our current experiences that there's kind of uneven inconsistent identification of what is a composite performance measure.

So, I'm going to stop there and see what your thoughts are in terms of being able to provide some (guidance).

(Alan Zavslavsky): This is (Alan). It seems to me that the essence of what we've been talking about in this committee – this task has been combining information that represents either multiple concepts and/or multiple measurement systems consist single measure and some of the things you've listed there do require special kind of attention but they aren't the kind of things that we've been discussing. So, in particular, the use of the (shrunk) shrinkage estimators with hierarchical modeling, we definitely want people to look at that who are familiar with the (existing) principles and so forth, but I don't think it's a composite measure in the terms we've been talking about it. It's certainly a single measure that a single concept – a single data source.

The kind of things that we've been talking about are really where, you know, you're putting together a survey measure and clinical measure or you're putting together several related concepts but about things that are measured by the same set of items and so I think we should restrict to this – to that kind of thing and leave the others to perhaps (inaudible) and some other (exercises).

(John Birkmeyer): I'd like to follow-up. This is John Birkmeyer. At our face-to-face meeting, we discussed how to consider shrinkage estimators in the context of composites and specifically we talked about the example of the measure that

was endorsed by NQF under a different category by the Leapfrog Group a couple of years ago that involved mortality rates shrunk to volume weighted a mean.

I think that's a key distinction in whether something (counts) it's own composite measure is whether there is two or more input measures. A one-dimensional single outcome measure that simply gets shrunk should not count as (own) composite measure. So if all you're doing was reliability adjusting measures (of) mortality to the overall population base mean.

I completely agree that that's not (old) composite. If you are consider, however, for example, both mortality and volume in a single measure that results in a new indistinct score well then I would argue that that does in fact count as (old) composite measure and even more so if there's multiple measures that are going into a hierarchical model being shrunk and ultimately (spitting) out a single score.

(Alan Zavslavsky): I agree with that.

(Elizabeth DeLong): This is (Liz). I agree those are my thoughts exactly that the shrinkage is really a methodology and not a definition of composite measure.

(Sherrie Kaplan): This is (Sherrie)...

(Elizabeth DeLong): ... technology...

(Sherrie Kaplan): ... adjustment is different from adding more things to estimate a (construct). So, I think we're all talking about adjusting single-item measures in various different ways but, you know, if you add more to the adjuster, who cares. That's an adjuster. It's not adding precision because you're trying to develop a complex construct, and for me, the first exception under their single performance measures based on patient score from a composite instrument or scale, I don't understand why that's not a composite measure.

I mean, just because you concocted a single score out of it, it's still a composite measure. The Framingham risk score under this criterion wouldn't be a composite measure and certainly is.

(Karen Pace): OK.

Male: Well, I mean, to get to (Sherrie's) point, again, it depends on whether the individual components of that Framingham score are treated as separate measures – a separate performance measures, and if they are, then the Framingham score becomes a composite. If they're not – if the scores only value is as a – is as a score not as a set of components then it's not a composite. I mean, by the NQF definition.

(Karen Pace): (Yes).

Male: Otherwise, every measure becomes a composite, you know, it becomes – it becomes impossible to draw a line.

Female: That was – I think (our) risk score (with) that example because it includes age and I don't think we can influence age too much but so that's more like an adjusted than anything else but say you were – say we're going to have a cardio – cardio metabolic risk factor scoring. You created it out of lipid levels and, you know, hemoglobin A1c in this and then you put it together as a composite. That certainly counts and just...

(Karen Pace): Yes.

Female: ... as you treated this as a single score doesn't make it not a composite...

(Karen Pace): No, right, right. So, that first line, we need to work on the wording but, again, it's a distinction between, you know, that (Patrick) has been talking about are the items trying to get at one concept or multiple concepts. So, I think, you know, we need to be more clear about how we describe that.

But going back to (John Birkmeyer's) point at – so would you see that the – because you shrunk back to an average for a volume group versus the overall average, you're saying that you would see that as a composite and the question is, it's not just about terminology but whether these new criteria applied to it.

So, you would see that the additional criteria primarily about (2d) where you have to have analysis of the component measures, the aggregation of waiting (roles) and that those things would apply equal – would apply to that type of performance measure versus obviously what you would have to, you know, do for any risk adjusted outcome measure.

(David Shahian): This is (Dave) Shahian. I think we may be having done a very slippery slope here because that the distinction that (John) I think correctly makes for some measures where the particular procedure, for example, has a demonstrable very strong volume effect would not be a reasonable distinction to make in situations where there is not a strong volume effect and shrinking to the overall mean versus a volume specific mean will have different implications depending on those two scenarios. I think that may be a fairly complex argument for some of these committees to follow.

John Birkmeyer: But, (Dave), this is – this is (John). I totally agree with your point, but I don't think it's relevant to the question about whether (account) is as of composite. But I think that the problem that you just described would should they're heavily on the adjudication process and whether that composite was ultimately endorsed by NQF if there was a scenario whereby, you know – whereby endorsers were proposing measures that were a roll up of, say, mortality and volume for a procedure for which volume simply didn't matter. There would be no demonstrable advantage of adding volume to that measure and sort of a thoughtful adjudicator wouldn't endorse that above and beyond the existing independent measure of mortality.

Female: So, what if you were doing some other kind of shrinkage to sum up their average character so by an average based on some other characteristics? So, what if you chose that you were going to get the average for all community hospitals versus academic medical centers and use that as your basis for the average for shrinkage? Would that constitute a composite then or...

Female: I still see that as methodology. It's one input that's being adjusted for the type of hospital. It's not being incorporated into a composite of other things that are related to performance.

(David Shahian): Yes, I mean, that sounds like a broken record, but it gets back to the argument of whether that factor that's being used in the shrinkage is itself interpretable as a performance measure. And...

Female: OK.

(David Shahian): ... it is the case that volume itself is used as a performance measure in some context and in fact NQF has endorsed some volume indicators as performance measures. I don't think that anybody would endorse hospital ownership as a performance measure.

Female: Right.

(David Shahian): So, if you're simply shrinking based on a hospital ownership, that would be in the adjustment category. If you're shrinking based on some structural characteristic of the hospital that is construed and interpreted as a performance measure, then that becomes a composite and it becomes important to show that compositing those two measures that using volume per shrinkage, as (John) has suggested, actually improves the reliability of the scores at the end.

(Alan Zavslavsky): Just to avoid confusion, what we're talking about here is not adjustment. Adjustment would be, for example, if you try to remove the effect of something such as patient risk factors and the prediction of mortality. But what we're talking about is treating the – basically treating the kind of the composite of a structural measure and then the outcome measure, which is, you know, balanced aim but it's definitely mixing together things that are conceptually different and...

Female: You're right.

(Alan Zavslavsky): ... different sources and I think that's clearly a composite and would have to be justified as people commented by showing that's same as entering the composite is actually a performance, you know, a sensible performance measure.

Female: However, some people do refer to that as the reliability adjustment.

(Alan Zavslavsky): Well, the reliability adjustment piece is – it has nothing to do with adjustment incentive of risk adjustment. It's...

Female: I know, I know, I know.

(Alan Zavslavsky): ... (the weight) from the composite.

Female: Right.

(Alan Zavslavsky): Based on the reliability of the two measures.

(Patrick Romano): Yes. Just to make a slightly more general comment here, so I would just encourage us to try to minimize the emphasis on additional analysis and evaluation. In other words, I think what we've been saying in both our in-person meeting and in discussions is that composites require different metrics, not necessarily additional metrics but they require different metrics to evaluate them.

So, I don't if there's a way of tweaking some of the language to recognize that, you know, specifically on the reliability domain, we're not – we're no longer concerned about the reliability of the individual components. Now, we're concerned about the reliability of how they work together as a composite. So, it's calling some of the composite shift the focus of attention, but it doesn't really add to the analytic or evaluate the (burden) to shift it.

Female: Well, I mean, according to the way we have this and have the discussions have gone is that we have the reliability and validity of the composite performance measure, but 2d is about justifying the components and the way it's put together as a composite. So, that it would be additional analysis that your – so, unless – I mean, if we...

(Patrick Romano): Right...

Female: ... if you're suggesting that we don't have – that all they have to do is the reliability and validity is the final performance measure that's different than the way we thought you all were going. So...

(Patrick Romano): Well, just to play this out a littler bit, so one of your examples here has to do with the multiple complications list. And so, this is an area that I know something about so, you know, we have a number of indicators (technical interference) a single indicator that includes a whole set of (IC-9-CM) code within it and we happen to call it postoperative hemorrhage or hematoma, for example, and by virtue of the title it includes both hemorrhage and hematoma. We have postoperative deep vein thrombosis or pulmonary embolism. By virtue of the title, it includes both deep vein thrombosis and pulmonary embolism. We labeled that as a single measure.

But as the panel reviews, it has to decide whether it makes sense for DVT and PE to be included in a single measure. For composite, it's really no different. If we didn't put DVT and PE together with postoperative hemorrhage and hematoma, now an expert panel has to decide whether it makes sense to put those two concepts together.

So, it's really – I think there's no bright line here. I don't – I don't see a bright line. Ultimately, it comes down to how the developer chooses to label the measures and to promote the use of the measures.

(Sherrie Kaplan): This is (Sherrie). I think I'm getting a little lost again on the principles. Because the principles are if you're measuring a complex construct and you're using more than one thing to measure that complex construct, because by it's nature it's complex, then you have to have a conceptual basis for doing that. You have to have your head screwed on when you're putting this thing. You cannot add up apples and airplanes, but I thought the purpose of this section would to say, OK, then you got to test whether you were right.

And, you know, you have to start with some conceptual basis and then you do have to do the additional testing to see if you were right. And if the things that collectively you think measures some more prior order abstract concept, actually do hang together and reflect that in some empirical testing. So, (Patrick), I agree with you but I think that at some level we're losing the plot here.

(Patrick Romano): Right. I'm not disagreeing with you. I'm just – I'm just saying that we should think about really just the same – it's the same fundamental criteria that are being applied. It's just that the criteria have to be applied in a somewhat different way when we label a measure as a composite (inaudible).

(Alan Zavslavsky): (Inaudible) in reaction to (Sherrie's) comment. I get back to the (inaudible) that it's something that's done within a single – it's done in a single concept within a single methodology then you don't need to bring the composite measure process to there, because you'll have people of expertise on that concept and methodology on a regular review. But if it's more than one concept that is being combined or more than one methodology is being used that are (inaudible) different by combining the structure and outcome in the composite then the composite process (inaudible) it is that helpful. I think that covers the situation we've been talking about.

So, (Sherrie) does it, you know...

(Sherrie Kaplan): I wasn't sure I was clear on that. Yes, I don't know because you kind of – we're talking all over the place here in abstractions and when you get down to the actual example, which is what I thought that the NQF is kind of trying to plan for so that they could screen for things that, you know, you fix something up and with the set of criteria you could say, "OK, this is or this is not the composite."

It's very difficult for me to see the difference, (Alan), between, you know, creating a multi-item measure of the same concept that's, however, narrowly it's defined versus some (mover) construct that combines structure process and outcome and some, you know, overall quality marker. The principles are the same. You still have to test whether or not those things that you collectively said reflects some abstract concept actually do that and it's somewhat different for when you're combining these (mover) constructs into single composite score. But the process you then pick some different method for testing it, like interclass correlation or something like that for reliability, but you'd still have to go through the same process, right?

(Alan Zavslavsky): Well, I think we're trying to figure out – also, what we're doing here is directing an application to one or another set of people to look at it with a set of...

Female: Well, yes, I don't think it's necessarily that. I mean, I think, you know, definitely we recognize that, you know, any complex measure whether it's an outcome measure with risk adjustment or this composite measure requires some people with expertise to review them and that's an issue that NQF has to work out in terms of (feeding) their committees.

But the issue, you know, let's look at page 14. We added to, you know – definitely, we said that the NQF criteria applied to composite performance measures, but we identified two additional criteria, 1d and 2d, that would have to be met if it was also a composite performance measure. And I think what we're saying is we need to identify and this criteria will require some additional questions on the submission form in order to apply these criteria. And so, we need to know which measures fall into this category that would require this additional information.

1b is more of the conceptual and probably is not that big a deal. But I think, as (Sherrie) was saying that this is to present the conceptual and then 2d gets into empirical analysis of that conceptual model. However, you just define the quality construct and the rationale. So, you know, unless – yes, go ahead.

(Patrick Romano): Let me try asking a very practical question and see what the group thinks.

Female: OK.

(Patrick Romano): Let's take postoperative complication, which is one of the – one of the examples that you (mentioned).

Female: Right.

(Patrick Romano): So, CMS as proposed measures of postoperative complications, like for total hip and knee surgery, I think, that include a whole set of different types of complications; infectious complications, thrombotic complications...

Female: Right.

(Patrick Romano): ... cardiovascular complications and so forth. And they're all put into one list so they call it a single measure of postoperative complications after total hip surgery or total knee surgery.

Female: Right.

(Patrick Romano): By contrast, AHRQ has developed separate complication measures for postoperative DVT-PE, postoperative hemorrhage, postoperative infection, so forth and then it said, "OK, well, we're going to put these measures into a composite." We're going to measure each of this at the hospital level and then we're going to – well, at the patient and then (rate it) at the hospital level and then we're going to construct the composite from these components. But they may get to fairly (propose) approximation at the same place.

So, do those get treated differently? Should those get treated differently in the NQF process, because CMS has chosen to label it's measure as a single-performance measure of postoperative complication whereas AHRQ has chosen to label it's measure as a composite of multiple separate performance measures?

Female: So, that's precisely the question we're asking.

(Patrick Romano): So, what could anybody think?

(Sherrie Kaplan): This is (Sherrie). So, in one case, I understand better about what's going on. In one case, it's an index, right? You're just adding up a bunch of zero-ones, either had them or you didn't and so you're creating an index, a performance index.

(Patrick Romano): Right. So each patient ends up as a zero or one. If they had any of the complications on the list, they end up with the one.

(Sherrie Kaplan): Yes. So, they are one and so – and then at the hospital level you're adding up all of the scores, right, location scores. So, you can have – you get a one if

you're at the hospital if the patient had a DVT or some postoperative hemorrhage or whatever else happened to them.

(Patrick Romano): That's right.

(Sherrie Kaplan): So you can get the one in a variety of different ways.

(Patrick Romano): Yes.

(Sherrie Kaplan): You're getting a single score off of an index and the other one I'm confused about how the other one happens. Is it an index in the same way or is it somehow scaled?

(Patrick Romano): You know, in the other one each complication, each type of complication, let's say there are half a dozen different types, gets measured separately and each patient ends up with the zero or one score on that particular type of complication. And then those scores get thumbed at the hospital level and then the hospital gets a composite score, it's a weighted composite based on the six different components.

Male: So, (Patrick), the – I'm trying to decide though whether I think in either circumstance all of these individual complications are, quote, "performance measures" or whether you're talking about the AHRQ composite as you're describing it or the CMS performance measure that looks as a least complication. I'm just struggling to say that in either those of those circumstances those are performance measures, they're accounts of event and they're often termed, you know, patient safety indicators or quality indicators or, but I'm not convinced they represent performance measures that would individually stand on their own.

(Patrick Romano): Well, remember that they don't individually have to end on their own in terms of the reliability criteria, for example. But, you know, I mean, whether this is a discussion that obviously can be held within individual steering committee, but the way AHRQ has defined it's composite, it is a composite of individual performance measures.

So, I just – just want to make sure that we're all OK with this distinction that that the CMS measure is a single measure because of the single list of complication and it scored zero-one for each patient whereas the AHRQ approach is a composite because it is bringing together six different performance scores on different component measures.

(Sherrie Kaplan): This is (Sherrie). The NASDAQ index is a composite, they call the composite. And I think we're getting into sort of semantic, you know, details here that may or may not be ultimately helpful to NQF. But I think if I were looking at this I'd say this is not that much different from what NASDAQ does so let's, you know, let's call both of them the same and subject them to the same issues.

If you're going to add up all of these different things and call them complications, you know, undesirable complications from whatever and you're going to call that a quality indicator then it's the same thing as adding up zero-one, did you get an annual eye exam or annual foot exam and all that stuff. So, for me those two things are both of them are composite.

(Patrick Romano): Yes, well, that's exactly my question.

Female: And I see it. I see the opposite. I really think that one observation per patient is a single performance measure whereas when you look at several observations per patient and then combine them, you've got a composite.

Female: But we've already been a little – and there are several areas of inconsistency, because when we talk about on the positive side and doing things like the all or non-measures are one observation per patient and we've said that those would be composites.

Female: Oh, we did? Sorry, I...

Female: Yes. Well, I mean, this is, you know – this has raised questions on multiple levels so, you know, that it so, you know, there's...

(Patrick Romano): There's other space argument that the individual components are performance measures themselves.

Female: Right. Right. And I think it's not some – again, it may not be an issue if, you know, in (Patrick's) example if one was considered a composite and the other not where we would really run into problems if CMS has a list of postop comps for total hip and knee and submit them as a single measure whereas another developer doing a list of complications for some other procedure and submits them as a composite.

I mean, so, we need to have some consistency if really want them to do whether we call them additional analysis or work this into the criteria. And perhaps we need to go with the simplest definition where we have the, you know, two or more performance measures or two or more measures or two or more concepts I don't know how you wanted to do that regardless of how it's combined whether it's at the patient level with one observation per patient or, you know, computed performance measures. But we need, you know, some guidance because that's one of the things that, you know, obviously there is still question about as were talking about this.

(Sherrie Kaplan): This is (Sherrie). What's the dark side of being erring on the simplicity so you're going to get pushback from the measures developers?

Female: Potentially, but I think there's probably – I think there's probably more dark side to having inconsistency continue. So, you know – but yes, that could be – that's why I keep asking about the additional analysis. You know, if we've mischaracterized what goes into 1d and 2d as things that would be additional information submitted for composite then that's another thing we would have to revisit. But the way we've been reading this and thinking about it is that there wouldn't be some additional information and analysis that would accompany a composite, so that could be some pushback there; yes.

(John Birkmeyer): This is (John Birkmeyer). On one hand, I agree with (Sherrie's) point about question at least about what the harm is, about over inclusiveness. But certainly from a practical level, I would imagine that NQF would want to make sure that, you know, that it's (methodologic) expertise in composite measures like it was optimally concentrated on measures that really needed their input and certainly some need them more than others.

Let's say just to add a gut instinct level, I have trouble dealing as composite measures those that, you know, count as a one at the patient level if any of a list of things have occurred and I think the (easiest) example is the surgical complication case, you know. Certainly, you know, it's not a composite measure, you know, if you get quoted as a complication if you had one of 20 different things happened that's the way that, you know, morbidity has traditionally been defined in surgery and nobody has ever called it composite measure.

I would view a composite measure instead as one in which or a (proposer) in which a developer was putting forward a measure that was some new surgical monthly index where there was not only a list of complications that, you know, that were considered at the patient level but that there was like some waiting or some aggregation that was more innovative and simply presence or absence. And I appreciate that this is kind of a slippery slope and that the surgical complication example is simply just the flipside of the all or none question for process of care measures. But nonetheless, it just doesn't have a lot of traction here.

(Patrick Romano): Yes, I would agree and I suppose this is a devil-fabricated question. But I would suggest that may be the way through this is that the CMS measure is going to be evaluated by clinicians according to the clinical concepts of whether these are all complications that ought to be countered as complications, whereas the AHRQ composite has to be evaluated from the analytic perspective of whether it analytically makes sense to put together the six different component scores into a single score.

So, from that standpoint, I think I agree with (Liz) that there is a distinction and that the focus in the CMS case is going to be sort of entirely on the clinical concept and whether these things hang together, whereas in the AHRQ case there really has to be an analytic component of whether the pieces hang together analytically.

I don't know if that – if that helps at all.

Male: Well, I tend to agree with you on that, (Patrick). I just have real concern about being too broad here, because NQF would then be faced with every single survey tool, every single measure that rolled up a group of complications or single events that weren't necessarily distinct performance metrics would then be counted as a composite and I think that would be overwhelming for NQF and would get a lot of pushback from the major development communities.

Female: And I – I'm in agreement with you there. I do think that if we're going to require additional analysis for composites, we have to be somewhat restrictive in what we call a composite.

Female: Was that (Sherrie) (inaudible).

Female: I'm not sure you and I are in total agreement.

Female: Actually, yes, I am still struggling with the distinction. Because I really think that in the case where you're adding up a bunch of things that you say reflect surgical complications, absolutely the first thing you must do is make sure they make sense to be experts that those are in fact reflecting a cluster of things that collectively reflect complications as the result of surgery than when you go adding them all up to reflect performance of hospital you better test whether or not they actually distinguish one hospital from another on some dimension you call quality of care.

So, I still think that you should hold everybody to that standard and I don't think you're going to get long so I think that people who do – who create these collectives different from people who made individual measures are not going to flip out over this. But, you know, that's NQF's call.

Female: OK. We'll have to digest that a little bit more, I think, and see if we can summarize your discussion and come to something that we can at least try in terms of moving forward. I hear your, you know, the discussion about the similarities and the differences and which result in the best approach forward.

(Patrick Romano): Yes. But I think what all of us are saying is that it is hard to draw a bright line between composite measures and non-composite measures. And therefore, it's

the who's us in general to emphasize the choosing the appropriate metrics for value purposes and perhaps try to avoid getting too hang up over whether something is defined as a composite or not. I know that's difficult, but that's the challenge.

Female:

OK. May be what we can do is I'm going to diverge a little bit, because I really – I think we will help but may be as we think for this is to specifically look at 1d and 2d. This really would be the additional areas for composite performance measures and we need to make sure we have that right and then we can think of, you know, it may help as we're thinking about some of the examples we've been talking about may be that will help us distinguish.

But we thought for 1d, which is that for composite performance measures that we wanted a clear and logical statement of the quality construct, which would include the (representativeness) of the component measures and the relationship of the component measures to the composite and to each other. The second was the rationale for constructing a composite measure including how it provides distinctive or attitude value over the component measures individually and thirdly, how the aggregations of waiting of the component measures are consistent within representative of the stated quality construct and rationale.

So, we had one question from the TEP member whether we really needed the second item, the rationale, though I think this is where the whole decision-making context might come into play here, or at least until we have it written up into text.

So, a couple of things, it's first of all, what of these do we need for our composite performance measures? And secondly, our understanding is that (to) really evaluative component of – in terms of when a steering committee would look at this, is really – only that is clearly stated and logical, but this would then fall in – would get (2d) is where the empirical analysis related to this (command). So, I guess the first thing is, is this correct? That we would want a description of the quality construct, the rationale, and how the aggregation and weighting are consistent. So, I'll stop there.

Male: Personally, I like it. I think, two is just really an add-on to one. I mean, it's really just further explanation and elaboration of one. So, I think it makes sense, but...

(Dale Bratzler): Yes, and this is (Dale). I agree. I thought it was reasonable.

(Karen Pace): OK.

Male: In a way, it's just a restatement of things that you would take for granted in a single narrow measure, but that someone might fail to do with a poor composite measure, so it's reiterating the importance of something we already need.

(Karen Pace): So, (Liz), I think you had a question about number two, the rationale, and whether that was necessary, but it would – maybe you could tell us a little bit more, whether you thought that was just part of the quality construct, and we didn't need to elaborate it, or whether we should pull it out as the idea.

(Elizabeth DeLong): Well, I'm afraid I'm sort of flying solo here and that I'm probably the only person, who doesn't think that the quality of care basically overrides the audience for which this is intended. I kind of see quality of care as quality of care, and if it's from the patient's perspective or the physician's perspective or the payer's perspective, it still should reflect some level of quality and that's where I'm coming from.

(Karen Pace): OK. Yes, and that gets into the (whole of) decision-making context that we talked about.

Male: But I don't think that alters the concept that there should be a particular rationale for constructing the composites and demonstrating it's value.

(Elizabeth DeLong): Well, in my realm, the value is whether it leads to better outcomes.

Male: Well, that would be an acceptable statement under number two.

(Elizabeth DeLong): I mean, better outcomes – better outcomes regardless of who is paying or who is watching.

Male: You know, and I – I'm always afraid to say anything here, but, you know, so I agree in concept that we want better outcomes with composite measures that seems to make a lot of sense. I think for a lot of composite measures probably many that are in use, there's not a lot of great evidence that you can demonstrate and prove outcomes. So you take a group of individual measures for which there's good trial evidence that doing this is the right thing to do for patients. So that seems to make sense, but if you roll them up and do them all well, that you should see better patient outcomes.

But I just think that sometimes that's very difficult to demonstrate either from a review of the literature or in any short-term pilot of these performance metrics, because you're rolling up these composites in the real world, not in a clinical trial, and requiring that evidence makes it a very difficult thing to prove sometimes, so I...

(Elizabeth DeLong): I absolutely agree with you. Basically, what I'm saying is that the goal is better outcomes.

Male: Sure.

(Elizabeth DeLong): I think the demonstration of better outcomes is something we've discussed in another section and that is more tenuous, but the goal is always better outcomes.

Male: Right.

Male: Well, the goal is not always limited to better outcomes though. I mean, we do – I mean, I want to get into a broader philosophical issue, but we do believe that patients should have a positive experience in the health care system, even if that cannot be directly demonstrated to lead to better outcomes. We think that it's likely to do so, and we think that we are in a service industry, where it's important for people to have a good experience with care, so there may be other objectives of the measurement enterprise besides simply improving outcomes.

(Elizabeth DeLong): Yes, I could talk a while on that one. I do think that's analogous to students writing their professors, but I'll like to have that.

Male: I think the issue here is more of what's the rationale, you know, and why we have NQF that has better outcomes, but that doesn't mean that every measure that concerns things that NQF with that, it makes sense at a way looking the information. So, you know, you could put together a composite of diabetic foot exams, appropriate antibiotics for people who come in with respiratory infections and proper preparation of anesthesia. We wouldn't call that a good composite, even though they're all things that contribute to better outcomes, because they don't correspond to an appropriate routine of smaller concepts into a larger concept that provides a value to people looking at the information and the kind of composites.

We dealt upon at things that have a more specific rationale, like putting together things that are responsibilities of the same party and health care or putting together things that affect outcomes with particular type of patients or things, to get the things that affect these different types of outcome, something that provides some more specific rationales and the general one for that points to any measure we might construct. And as I think, like two is getting that.

Male: Yes, I think that's nicely described, and it really points us to what analysis reviewers will be looking for in evaluating composites.

(Karen Pace): OK. So let's then move on to (2d), because this is where the additional analysis and evaluation come in because we've talked about reliability and validity, we need to focus on the performance measure, and we – I know there were some questions about validity that we can come back to, but I want to stick with this composite analysis for the moment.

So, (2d) again would be something that would be inserted that was specific for composite performance measures and the first element of that is that the component measures fit the quality construct and add value to the overall composite. Several of you suggested that we take out the necessary, and we can talk about whether we still need to talk about achieving the related objective parsimony, but I think the question is here, you know, should there be analysis that demonstrates component measures fit the quality construct that's described under (1d)?

And the second element was that the aggregation of analysis that demonstrates the aggregation and weighting rules are consistent with the quality construct and rationale. And the third is about the extent of missing data, because when you get into these complex measures, that can be problematic.

One of the things, that came up in discussion with the CSAC and, you know, in the – maybe in some of their comments, is about, what are some examples of what this analysis would be? And so again, we – in the guidance column, we tried to pull some things that were already in the check into example, but I think we want to start first with, you know, are these things that need to be demonstrated for a composite performance measure or not? And if so, then we can look at the examples.

Male: That's unusual, we're speechless.

(Karen Pace): Well, let – yes, so...

Male: You can go on.

(Karen Pace): So if the component – so maybe it would help to talk about examples. So do the component measures fit the quality construct? So we talked about if the components are correlated, then analysis might be based on shared variations, such as factor analysis, items total correlation and inter-item correlation that really indicates these things hang together. If it's not – if the components are not correlated, then analysis demonstrating the empirical contribution of each component to the composite, or might be something that could be considered, or clinical justifications, such as correlation of the individual component measures to a common outcome measure.

I think it's a little less clear and, you know, a little more problematic to think about what those analysis would be if the component measures are not based on that underlying correlation and – but I think if – you know, in general if we're going to say composite measures need to, you know, defend their construction, that we have to talk about both component and the weighting. And I'll just say, you know, we've skipped over it, but we've already addressed that each component should meet the evidence criterion, which is the clinical evidence, and why you would even be considering that. This

really is getting at the empirical analysis that these things do work together or should be in the component – in the composite, sorry.

(Sherrie Kaplan): This is (Sherrie). One of the – one of the things that's missing here for me is some of these things, you know, were kind of following the – if you will – if you will, the sort of psychometric language in that – in that sidebar there. Whereas like interclass correlations, do these – do hospitals behave consistently across patients with respect to quality on these indicators and different from hospitals who behave differently from them and are also consistent across patients within their kind of constituents?

Female: Right. So we – that would be under our general reliability...

(Sherrie Kaplan): OK. Right. But it's not...

Female: ... in analysis.

(Sherrie Kaplan): Right.

Female: And so...

(Sherrie Kaplan): But it's not an item in total correlation, which is in that same realm, same cycle of nutrition. So for me, what the weighting thing should demonstrate is the value of weighting should add to the discrimination between units that are being compared, and if it does not, then there's – then the rationale has really stuck with whatever conceptual business you have going, but if it's not helping you discriminate one unit being compared from another, then the weighting is not adding value.

Male: It could be based on criterion outcome measure.

(Sherrie Kaplan): True. I mean, either if there is some grounded conceptual basis for doing and just because in your particular empirical sample that you didn't sort of demonstrate the value of weighting and you would in the more robust sample or whatever, that's one concern and you'd certainly want to keep that in your considerations when you're reviewing the measure. On the other hand, if it

consistently doesn't demonstrate any value across multiple different samples you've tested, then the weighting scheme may need to be revisited.

Female: Right. So I guess one way to or another way to ask this question is, "If we really hold to and require that composite performance measures, is it reliability analysis?" For example, signal to noise, and actually do empirical validity testing, do we need any of (QD)? And, you know, or do we expand reliability and validity to include some sensitivity analysis about different weighting schemes?

Male: The part to get all of the possible rationales summarized in a – in bullet point like that, then it maybe better to leave them a little general and then have some discussion in texts of some of the different rationales that people might have, or weighting schemes and – or putting together a group of measures in general.

Female: I agree, but that would be under the rationale – the quality construct and rationale, so our question is – you know, I think everyone agrees that it does need to be clearly described, but then is there any additional analysis that has to go with that, or is it just – to just going to go to come down to showing that the ultimate performance score is reliable and valid, talking about reliability in terms of the signal to noise and that you can actually see differences between providers and, you know, some kind of validity testing.

If that would be sufficient and that simplifies things, and it takes away some of the issues we talked about of classified – you know, some of the issues we've talked about with whether things are classified is composite or not, because it more clearly fits into our general criteria with the caveats that you really do have to focus on the performance measure, not just, you know, data element reliability like (interradial) reliability.

Male: Well, I've – you know, we've said there are several different rationales, but not always were necessarily going to present, or why you would create a composite at each of those rationales would have a corresponding that of empirical analysis. But if your rationale is that these are things that are really closely correlated with each other and can well be summarized without these

as mentioned, then the analysis is a correlational analysis. The rationale is that these are all things that are predicted as a particular outcome, and the evidence, either your own data or from the literature on those predictive relationships, would be the empirical analysis and so forth.

Female: OK. So that – is that kind of speaks to that we do need some additional analysis that correlate with those things in (1d), the rationale and the quality construct?

Male: Yes. But this, I think, it's impossible to summarize them in a bullet point. You know, that maybe in the text, we could talk about some of the different rationales by creating quality measures and corresponding empirical analysis to back them up.

Female: OK. So I guess then the question is, are the – you know, for criteria, we do have to have the somewhat concise, so are the things under (2d), are those general enough that they encompass what we would be expecting without, you know, getting into the specifics, or does the wording need to be changed in some way? Do we want to know that component measures fit the quality construct that there are some analysis that the aggregation and weighting rules fit the construct and missing data? Are those things that we would want somehow addressed even if we can't give all of the specifics of how they should do that?

(John Birkmeyer): This is (John Birkmeyer). I think what you have is pretty good there. I don't think it's overly obtuse and I think each one of those three criteria are worth putting in there. If you want to make it more concise, for me, the first thing to go would be the missing data piece that's important, but that seems more of a (tree's) issue than a (forest's) issue.

(Sherrie Kaplan): This is (Sherrie). Having a veteran of the missing data problem and composite measures, that's one that actually can really trip you up fast, so I would recommend you leave that one in. The rest of them, to me, look general enough, so that if somebody can't figure out, you know, how to do that and apply it, can't they always call you or, you know, ask – is there – is this is an all or when we treat, is there a help desk at NQF, if somebody gets...

Female: We can and I guess what we could do, you know, we've talked about trying the – over the next several months, tried to build out that appendix B with some examples and maybe that's where we need to, you know, for now, not try to put it in guidance, but leave it for trying to build examples. But yes, they can contact us and we can point them to some relevant expertise, but I think, you know, I guess for now, we need to make sure that we have the criteria wording correct, if people can live with the one, two and three under (2d).

Male: I think it's reasonable.

Female: OK. So having talked about (1d) and (2d), I'll just to see if anything else has occurred to you about our discussion about classifying things, because that's the intent, that things that are classified as composite would have to meet those two additional criteria; (1d) elements one, two and three, and (2d) elements one, two and three. And so, you know, there's – we don't want – we're not trying to corner to market of the term composite that's long gone, but in terms of applying these criteria, that's what we need to be concerned about correct classification and that we're treating everything equally.

All right. Well, we'll like to mold that over a little bit after the call. And maybe we will take a few minutes to address a couple of other key questions and then we will have a time for some public comments, and then we'll decide our next step. But I think, related to what we've been talking about is we were asked to come up with a definition of quality construct and we've put that on page – where is it? Page nine. And again, this is kind of new for you to take a look at, so you may not want to talk about it right now. And it's also in the glossary, but it would be the same definition, because we've talked about a definition of a quality construct and one of the questions was, "What do we mean by quality construct?"

So I'll just point that out to you, but what we have down here is the quality construct describes the overall area of quality, for example, quality of CABG surgery – the components of quality, for example, preop data blockade, CABG using an internal mammary artery, CABG-risk adjusted mortality – the conceptual relationships between each component and the composite – and

then the relationships among the component measures. And then obviously, if we had appendix B done, that would leave them to some examples that we need to build out. But if you have some thoughts now, you can speak out or, you know, maybe mold these over, and send us your suggestions of modifications or striking or adding things.

OK. The other next thing, I guess in terms of reliability testing and this could also go along with our conversations about what is and isn't a composite performance measure, is that our discussion and what we've put in to the guidance in the table with related to (QA2) reliability – our general reliability criteria is that reliability should really focus on the performance measure score, not reliability of individual data element or into each role performance measures that are included. And there's a couple of questions here just to confirm that this also applies to all or none, or any none – any or none composite measures as we've been talking about them.

And the second thing is that we have had under both reliability and validity about the component measures that if they were going to be disaggregated for in accountability applications, then reliability and validity would need to be demonstrated. We had comments from the test and from the public commenter's of, you know, "How is this going to be known?" – you know, the use of the measures is outside of NQF control, and what were suggesting is striking that and making clear later on under usability and use that NQF endorsement of the composite measure is for the composite measure, and it doesn't extend to the component measures, unless they were individually submitted and reviewed, and endorsed as individual performance measures. So we think it would just be cleaner to have just focused on the composite performance measure and wanted to see if there were any objections or additional thoughts about that.

(Sherrie Kaplan): This is (Sherrie). I think it's a little off track, but...

Female: OK.

(Sherrie Kaplan): About the scoring transparency of composites, especially for proprietary measures, is there going to be a way in here somewhere in this document to

underscore the importance of transparency of scoring? Because it was – it dawned on me that some of these folks may have – you know, proprietary folks may have a measure that they really don't want to reveal the sort of underpinnings of these. Is there an imperative in here or is that not your intent to have the scoring to be absolutely transparent?

(Karen Pace): That's a good question and I'll tell you in terms of NQF and proprietary measures that if someone submits a proprietary measure or component – a proprietary component to a measure, it has to be totally accessible for the steering committee to evaluate, but after that, if it would be endorsed, you know, they can have fees or licensing requirements regarding the proprietary component.

Now, we can say, you know, I don't know to what detail you're talking about, I mean, certainly, you know, we can say that at a minimum, you know, obviously we wouldn't want to say name of a composite without people having some sense of what's included in the composite.

I think that would be totally useless to most people, but in terms of like, you know, if you were talking about weighting of risk adjustment model, you know, that was proprietary. It has to be available for review (inaudible) after endorsement. It's not that it has to be publicly available, like in, you know, it could have some requirements to it.

(David Shahian): This is (Dave). Just sort of – just an off the top of my head comment that I think if one submits a measure that's to be used for public reporting and evaluating the performance of providers that there is almost an ethical responsibility to have all the technical aspects of that measure available to the public.

I've taken that stand, so I've written editorials about it, and every measure that's come out of the (STS), you can see every intercept, every coefficient, and I really think that's a stand that NQF should adopt, but it's kind of off topic.

Elizabeth Delong: I don't think it's off. But I think it was good that (Sherrie) brought it up, and I absolutely agree with you, (Dave). This is (Liz).

(Karen Pace): And I think that we can certainly put that in as a recommendation that would need to be addressed by the CSAC, and ultimately, the board, because I think, you know, it definitely has a broader implication, but it's something that certainly is highlighted when you get into the complexity of composite performance measure.

So, I think that's certainly something that we can put in here as a recommendation that would need further review and exploration by the CSAC and board, but certainly can be put in to the recommendations.

Is there anyone that would want to speak to the other side of that or against that recommendation?

OK.

(David Shahian): This is (Dave) again. Can I go back – I was on mute for a while and you...

(Karen Pace): OK.

(David Shahian): ... brought up the question of quality construct, and I just wanted to get some clarification...

(Karen Pace): Yes.

(David Shahian): ... maybe ask a statistician to weigh in, but when I think of a – I guess when I think of a quality construct, I am thinking of the underlying unobservable latent construct, which is estimated using observable components of that construct. That's a different flavor that I get from reading this paragraph.

(Karen Pace): OK.

(David Shahian): So, I'm just wondering, you know, again, ask a statistician, how they feel about that.

(Karen Pace): Great! Would you suggest then that's all that needs to be described, because when she wants some description of what people are considering...

(David Shahian): Yes, but I think it's all lumped together here, so I might – I might have – again, I'm asking the statistician...

(Karen Pace): OK.

(David Shahian): ... but I might have an introductory sense about if in fact they agree, an introductory sense about we're measuring an unobservable, and then, how do we measure that, will we measure that...

(Karen Pace): OK.

(David Shahian): ... by individual using individual components, estimating what kind of relationships they have and so forth. I mean, is that – am I off track there?

Male: Well, I'm not statistician, but I...

(Elizabeth DeLong): With that I'm still looking in the document for what paragraph you're on.

(Karen Pace): OK. It's on page nine under that quality construct and rationale header.

Male: These lines gets opened if we ask, right?

(Karen Pace): OK.

Male: I think we were getting a broader concept of construct, you know, trying to avoid the clinimetric-psychometric distinction...

(Karen Pace): Right.

Male: ... by recognizing that a quality construct does not necessarily have to be an unobservable latent construct in a way that...

(Karen Pace): Right.

Male: ...defined in some context.

Male: That's fine, as long as – yes, that's fine. I think it may – it may be confusing for some people who are used to thinking in psychometric terms, but that's fine. Thank you for clarifying.

(Karen Pace): But certainly, you know, when you get to take a look at that and see if there is some way to make it more clear that we've captured what we want people to actually describe.

(Sherrie Kaplan): (Alan), this is (Sherrie). Remember when we were having a little dialogue a while back about the point of, you know, creating a composite measure to, you know, reflect a more complex construct like quality, and I would argue quality is a latent construct. There is no measure of quality.

(Karen Pace): Right.

(Sherrie Kaplan): I mean, that's the (obstruction) we all made up. But the point of doing – creating a composite to reflect quality is to reflect the complex nature of it, that it's multi-dimensional, it has multiple cards, and the best we can do to serve the health service delivery community in evaluating quality is reflect that in the things we use to measure it within sort of the sub clusters of quality that are uniquely important in those change over time to the health care delivery enterprise.

So, you know, so it is – for example, you wouldn't use all obstetric measures to reflect quality of care because they reflect quality of care for various, you know, a subcomponent of it. Same with pediatrics, you wouldn't use, you know, immunizations to reflect quality of care, because that's one important part of it, but it's not the whole banana.

So, you know, this piece, I think, we – I sent (Alan) some language back, a while back, about maybe putting in here the point of reflecting quality of care using component – composite measures versus individual measure is to actually give a more accurate reflection or a better reflection of the complex nature of quality of care.

(Karen Pace): And we do have that in here somewhere, (Sherrie), and maybe we can, you know, we may need to pull that into this section as well, so we'll take a look at that.

(Sherrie Kaplan): Yes, I remember reading that somewhere.

(Karen Pace): Yes. OK.

(Alan Zavslavsky): Yes, and I think the only thing I picked up on a little bit there is when we talk about maintenance composite measure, I understand the positive compositing. It doesn't necessarily give us a definition of what is a composite measure, and going from a specific to the more general summary is compositing, and I think we can describe that process pretty unambiguously.

It doesn't answer the question of, you know, when does it become a composite other than if you had a bunch of measures already and you put them together, you know that...

(Karen Pace): Yes.

(Alan Zavslavsky): ...as a composite. I don't want to get hang up on that anymore, I – and as just suggested both, you know, on the process rather than the outcome what you said to measure development is in some ways is the center that bind.

(Karen Pace): OK. Should we do...

(Patrick Romano): (Karen), I am going to have to...

(Karen Pace): Yes.

(Patrick Romano): ... step off a few minutes early, I'm sorry.

(Karen Pace): OK.

(Patrick Romano): So, maybe if you could orient us to sort of in general terms, whether you've responded to all of the comments that came in during the comment period, or whether there are particular things that we should attend to in our own consideration after the call.

(Karen Pace): Right. So, in the Excel file, and we also put it into a Word document, depending on, you know, your preference for reviewing materials, all of the comments, and we did have draft responses for your consideration, and in the Excel file, we actually highlighted the ones that are some of the things we've

already talked about, but really that we were targeted that most interested and having some of the TEP's input, so those are highlighted in the Excel file in yellow.

And so, we, you know, invite you to take a look at that. We tried to address most things in the – in the, you know, draft report with either questions in the comment balloons or in actual, you know, red lines. So, I think my sense of this call is that we're probably not ready to have a final document to take to CSAC, and that we need to really kind of make sure we have this language right around these issues that we talked about and get something out to you all with specific questions perhaps.

If you have any suggestions on path forward, we can do that, but that's how we identified particular comments. Do you have any thoughts, (Patrick), before you leave us?

(Patrick Romano): No. That sounds good. I just – I just wasn't sure, you know, the extent to which, you know, you had addressed, you know, every one of these comments or if there were others that required our consideration.

(Karen Pace): Right. So, we did an initial review and kind of a draft response for each one, but certainly subject to you all to review.

OK. Maybe, before we get into next steps with the TEP, operator, would you open the lines and see if there are any comments from the audience.

Operator: At this time, if you would like to ask a question or make a comment, please press star one.

Again, that's star one.

You have no questions.

(Karen Pace): OK. All right. So, with the TEP, we have a few more minutes left, so maybe we'll just tee up at least one more question or issue that we thought probably needed some resolution, and that's about validity testing. And if we go to the document, the table on page 15, it's (2b2) validity testing, again, our

understanding from the discussions and our initial review of the criteria with you was that validity testing for composite performance measure should be done for the overall composite, not the individual components.

However, we talked about that this may not always be feasible at the time of initial endorsement, and in that case, our understanding was that validity of the component measures must meet the NQF criteria and by endorsement maintenance, have validity of the testing of the composite performance measure, so that validity can be demonstrated.

So, I think we had a question from at least one TEP member about, you know, pressing for empirical validity testing, even by the time of endorsement maintenance, one or two comments about the issues of validity testing. And so, I just wanted to just bring that up for your review and just to confirm that this is kind of the way we should be progressing.

I think, some of the discussion we have that resulted in this language was the fact that, you know, how you construct the composite can affect the validity. And so, you know, we wanted to focus on the validity of the final composite performance measure rather than the components or just phase validity, but I'll stop there and see if anyone wants to suggest a change to this or reaffirm that this is the direction we should be moving.

Elizabeth Delong: This is (Liz), I like the way it's written.

(Karen Pace): OK.

(Dale Bratzler): Yes, this is (Dale). I agree.

(Karen Pace): OK. All right. Again, we'll follow up with you on this, because I know that as (Patrick) said at the outset, we know that – we know we've had to work over the holidays to get to step out to you, but that's a different expectation than expecting you to do so as well.

So, I think that we'll regroup here and maybe follow up with you with some specific questions, ask you to take a look at what's in the comment table as well as this latest draft and certainly feel free to send us anything that I think

will just need to digest some of the conversation we had, particularly about the classification and make sure that we all get on the same page with that for moving forward.

If any of you have any specific suggestions on what would be most useful to you, in terms of us, you know, following up with you, we'd love to hear that, because we want to make it as useful and sufficient as possible.

OK. Well, you know, if something does occur to you, we, you know, welcome your suggestions and input, whether it's on the specifics already there or suggestions on how to move forward, we'd love to hear that. We will probably kind of get out thoughts together and regroup to see if we think we need another conference call and whether we can handle some of these, you know, through sharing things through e-mail and online.

And I will stop there and see if any of you have any last comments or any other things you want to bring up that we haven't already brought up.

Female: So, what are the next steps again?

(Karen Pace): Well, I think, we need to regroup here a little bit to see what we think are the next steps. In general, the next steps are we have to finalize a document with the recommendations that will go to our consensus standards approval committee for approval. And so, we want to make sure that it reflects the consensus of this expert panel how best to accomplish that, and I think it's what we need to think about.

Obviously, you've had a lot of materials, and you know, we would ask you to review those and send us any of your feedback. If we have specific questions that we, you know, really want to nail down, we'll send those to you. But again, if you have any suggestions on how that would be most useful to follow up with you, we'd love to hear those. Thank you.

All right. And operator, you want to check one last time if there are any comments or questions from the audience?

Operator: Again, to ask questions, please press star one. You have no further questions.

(Karen Pace): OK. Well, I would like to thank the technical expert panel for all of your insights and questions and suggestions. And I'll thank you in advance for continuing to hang in there with us over the next few weeks so that we can get this right, and we'll follow up with you once we kind of regroup and have a plan on how to proceed, and you know, send us any suggestion.

OK. Well, thank you everyone.

Male: Bye.

Female: Bye.

Female: Bye.

Operator: Ladies and gentlemen, this does conclude today's conference call. You may now disconnect.

**END**