

NATIONAL QUALITY FORUM

Moderator: Elisa Munthali
February 21, 2013
3:00 p.m. ET

- (Karen): So one way we might proceed is if you think that we capture the main issues in the memo for discussion we could use that as the basis of our discussion today and we can see if there's anything that we missed that any of the TEP members would like to add. And I'll just stop there and ask Patrick and all of the TEP members, first of all, if there are any other issues that we didn't address in the memo or is that a viable approach to go through the call today.
- Patrick Romano: Hi. Yes, I think it is a viable approach. I think that there was still some open controversy perhaps in debate regarding the definition, and so I'm not sure if you were – it doesn't seem to have gotten into the summary memo, so I'm not sure how you are planning to handle that.
- (Karen): You mean the – well, we have the definition first and there seem to be pretty much agreement around the definition, the question came up about different types of composite measures and whether they would be considered a composite or not. Is that what you're referring to?
- Patrick Romano: Exactly.
- (Karen): Right, and I think that's at the end of the memo. You're right, I think that was the area with the most controversy and we thought may be we could get through some things more quickly early on, but we could certainly readjust. And I'll just mention that the way that we ended up handling that in the report is – and I think as we laid out in the memo is whether it was consensus or substantial consensus and agreement on what would be considered a composite. But that's what we would start with in those areas where there is

still controversy and lack of consensus, I guess our thought with that it's unlikely to happen in this call and that perhaps for now we'd leave those off and we visit them in the future.

But I think, you know, that certainly us for discussion. So, may be if we go to the ...

Patrick Romano: Right. The reason, (Karen), I would propose to advance that ...

(Karen): Yes.

Patrick Romano: ... is I'm actually going to offer a proposal based on my own review of the comments, the survey, and other discussions, it may be completely rejected. But anyway, depending on how we deal with these examples, it may simplify the rest of the discussion with respect to the specific criteria.

(Karen): OK, that's fine. So, we can go back, starts at the bottom of page four, the Identification of Composite Performance Measures, and the (Lisa), I think you've got it (have one) webinar.

(Lisa): Yes. I hope everyone can see it.

(Karen): Anyway, you can make that bigger. Or what can we do at this point to make it bigger on our own screen?

(Lisa): I'm going to try to make it bigger, so just bear with me for just a second.

(Karen): OK. And I think we each have a way on our own webinar to make it larger as well, OK.

(Lisa): Is that better?

(Karen): Yes.

(Lisa): OK.

(Karen): So, probably we need to go to the next page. Based on the results of the survey, we had agreement where I think we had agreement in past is that

certainly two or more individual performance measures that are combining to one score sort of provider to those agreements that were composite as well as the all-or-none and the flipside of that, a little less agreement but still seemed to be agreement that any-or-none is where we had agreements that those would be considered composite performance measures and the rest of that list than there with the group that we would not be classified as performance measures.

And keep in mind that this is for purposes of NQF endorsement in terms of submitting the measure, submitting information that would be needed to evaluate it and how it would be endorsed. So, I'll stop there Patrick and if you want to, you know, talk with us about what your suggestion is, we would like to hear that.

Patrick Romano: Sure. So, I think that where there were some issue that people were having trouble with is the all-or-none versus any-or-none and the all-or-none composite is kind of easier to see those as composites, because they usually come out of separate process measures that are developed as individual measures and then somebody says that you need to do all of these things and then it turns into an all-or-none, quote/unquote, "composite."

The any-or-none is a tougher problem, because the way that usually happens is that people identify 10 different types of complications or adverse outcomes and then they put it together and they call it a composite. And this is very classically done in clinical trials, where we often have outcomes that represent any of three or four different related outcomes.

So, I think that as I was thinking about these all-or-none and any-or-none measures and applying them against the criteria that we have discussed, I'd like to propose that instead of viewing these measures as composites, perhaps we should just use them as individual measures and that they should be submitted and reviewed first in just the same way as individuals are currently reviewed. Because if you take a part almost NQF measure, there's Boolean logic within than measure. And so, I'm having a little trouble with where we drag a line as to when that Boolean logic raises it to level of a composite.

Fundamentally, the performance metrics, the reliability metrics, the validity metrics for these composites should be, if we call them composite, they should be the same time of metrics that we would apply to an individual measure. So, I'm not sure in the case of this all-or-none or any-or-none I've just been wrestling with how the criteria apply and whether it's necessary to apply the criteria. Because as I go back and look at the traditional NQF endorsement criteria, I think all of those criteria can be applied to these any-or-none or all-or-none measures no differently from the way they would be applied to individual measures that don't happen to have that Boolean logic.

So, that's my proposal on the table to kind of reconcile that what seems to be an artificial distinction between these any-or-none and all-or-none and non-Boolean measures.

Sherrie Kaplan: Patrick, this is Sherrie. Now, I'm more confused than I was before. And for, you know – again, I think we all have to think in terms of measurement science and then coming from somebody trained in psychometrics and coming from that perspective. But if somebody who is now, right now teaching a course in measurement to clinicians, I know that there are, you know, formative risk as reflective types of measures.

Reflective types of measures are multiple items that reflect the same construct. Formative measures are ones that may be shouldn't even hang together, but they all cause some under what – some latent construct, so like, you know, losing your house and having your house burned down and having your wife divorced you and all that causes stress, but they may hopefully be related to each other, so that's formative. Reflective are mood disturbance and sleep disturbance and feeling sad and all of that stuff that reflects depression.

Anything with more than one element to measure in it that more than one measure is required to either or reflect or create a formative model to me is a composite. And I'm sorry I don't get the scoring of these measures if it's relevant, how their score is relevant. Anything that has more than one measure in it to create it is a composite. How it's done, how it's scored after that is to me irrelevant.

So, I think that we're getting hung up on the wrong thing. I think that, you know, when we got multiple things that we're going to stick with the formative model, OK. But still the scoring of that multi-item construct doesn't matter. What matters is this multi-item.

Patrick Romano: Well, but you see, it's a semantic argument whether it's multi-item or not. And that's the point that it's not – so, I would say that these kinds of things that people call composites aren't necessarily either formative or reflective.

In other words, people may say, "When I go into the hospital for a surgery, I don't want to have a heart attack, I don't want to die, I don't want to be ignored when I have pain." There may be no correlation between those three things, they may not – they may not be measures of any underlying construct, but they all reflect things that patient value.

And so, one may choose to put them together and you get into kind of semantic argument. So, if you're talking about complication measures, there are five different complications that we have and one person may say that it's a single measure because it's whether or not a person has complication. Another person may say that it's five measures because you can disentangle it into five different types of complications and fundamentally that comes down to a semantic argument.

So, that's why I'm trying to suggest a way to avoid that semantic argument and the words by saying that if you're constructing a measure based on a list of 5 or 10 items and you're calling this, you know, a measure of complications or a measure of adverse outcomes or whatever you want, that is not – that we can evaluate that just as we would evaluate any ordinary NQF measure. We can evaluate its reliability, its validity, so forth as a single measure.

Sherrie Kaplan: No, you see, first of all I disagree, it's not semantic. It's conceptual and that's very important as a conceptual argument. What you just described is a formative. (Albert Einstein) would agree it's a formative model for how those measures will go together. You wouldn't necessarily test them for how much they agree with each other because they can be very unique in the mortality argument, you know. You can't be dead and have complication.

I guess you could if you were dead from the complications, but you know what I mean? It's a formative model and there are ways you catch formative models if the assumption is formative. You wouldn't do, for example, a Chrome Box Alpha on those kinds of measures.

So, I think we're getting – I think there's a basic disagreement here that we should understand that one type of measure is formative, where the things are being added up in the index form and together they create – they represent a composite. But they wouldn't be tested in the same way of reflective model of the assumptions of a reflective model would. But a multidimensional reflective model is not – again, you would test that very differently from a formative model and the one you described.

Patrick Romano: OK, well, I guess – I guess we may just disagree. The argument that I'm trying to make it seems to be failing. But the argument to me is that whether – again, to get back to the example of complications, one person may view one user, one stakeholder may view complications as being a single measure. Another person or another stakeholder may view it as being 10 different measures that are put together with 10 different types of complications.

And that to me, I mean – I mean, yes, we can ask the developers to declare what their view is, but at the end of the day that is essentially an arbitrary choice and some developers may choose to say that it's a list of 10 that are scored any-or-none. Another developer might choose given the same design of the indicator to call it a single measure. So, that's why I think that we could simplify our task by just saying that these are simple measures if they're constructed using Boolean logic.

Sherrie Kaplan: I don't want to get into like a tangle on this, but I think NQF should ask the measures developers to declare whether or not they're assuming a formative or a reflective model as how they're proceeding. Because you're adding up things to create a single score that's how you're scoring items, not whether or not you have a multi-item measure that you're using to create a composite.

Again, in the example of losing your house, going to jail, being divorced, distressful life events-type thing, those things would not necessarily – they

would be added up to create because they cause stress the composite, but you would never consider them as a single item, a single measure. They're not – you're using multiple measures to evaluate a latent in construct. That latent construct is a single construct, but I'm having a hard time with what you're proposing at, I guess.

Patrick Romano: There are lots of single items out there in all measurement domains of whether a bad thing happened to someone. And again, the clinical trial field is full of those kinds of measures that are – where several different outcomes are put together into a single measure. And fundamentally, when we power clinical trials, when we design clinical trials with those measures, we treated as a single outcome.

Sherrie Kaplan: Yes, it's not whether they're a single construct. I think we're messing again with our language. There are single constructs but they're not a single item.

Jim Chase: Patrick? This is Jim Chase.

Patrick Romano: Yes, please. I'm done talking for a while.

Sherrie Kaplan: Thank you.

Jim Chase: I was – I'm trying to understand what you were proposing there, in the first bullet point, if I understand the first two in this paragraph. Then you're saying the second one may be we don't need to do it. I was getting a little caught on that, because I had been – again, putting this through the screen of hardware.

What we're trying to accomplish here is to give guidance to the measure stewards that are submitting this information about where they have to go through this extra steps and for the NQF itself about not, I mean, trying to minimize where we're capturing this where there's going to be extra work being done.

It sounds like you wanted to get – you're saying may be both all-or-none measures and any-or-none don't have to be included. I've been more on the side of thinking any-or-none measures is where it really gets foggy for me, because I think – I think I'm understanding your point.

I can see lots of practical examples where I can make it seamless if it's any-or-none, but it's really something where there's these multiple options of meeting that particular measure, like for a high lipid level, either the lipid is under control or you're on a statin.

To me that is not a composite that's just two ways to get to the same endpoint. Whereas putting together blood pressure control with the lipid control is a – and that could be a – that would generally be an all-or-none and that is a composite and I'd like to see the extra steps being done about what's the reliability when those are combined.

So, I guess, I was just expressing. I'm more with you if you were saying and I know there were some disagreements about this. But any-or-none, in my behalf for us or at least by the little guidance about, it doesn't rise to the level of a being classified as a composite if it's really something that is multiple ways to achieve the same outcome.

Karen Pace: This is Karen Pace. And I'd like to just, you know, to date NQF has had any-or-none measures that have typically been a list of complication. But sometimes that list might include readmission and mortality and then some other kinds of complications. And typically, you know, we have considered readmission, the mortality as individual measures even though in this case they might be just at the patient level.

And so, it gets a little tricky when, you know – and actually we haven't had any to date but you could conceive of any-or-none measures that would be profit measures where it could be things that would designate overuse or inappropriate care that the patient received any of these processes that would be overused or inappropriate.

And so, you know, to make a decision based on, you know, the one example we've had of any-or-none, which is the list of complication, it makes it a little difficult to move it to make that decision, you know, whether that's a composite or not based solely on what the topic is.

Nancy Dunton: This is Nancy. I am more comfortable in thinking about the definition of the composite if we actually go back to our original definition, which was that a composite was some kind of aggregation of multiple concepts and you have can have measures of concepts. And so, the any-or-none measure failed for me because – as well as the all-or-none quite frankly, because it's one content such as prevention.

And, you know, some of this semantic. But if we ask the developers to respond to guidance on composites by having them first discuss their conceptual model that involves more than one component or concept then I think it's helpful.

Patrick Romano: Thank you, Nancy. That I think clearly expresses – more clearly expresses what I was trying to say. And I think what has get in to the criteria is that – is that we have proposed criteria that focus on disentangling the composite and establishing the reliability or the validity of individual components of the composite. But if ...

(Karen): No, no, no. Patrick, can I (cut) you out. The way we've written this up and I thought that (inaudible) agreed that we aren't focused on the reliability of the individual components that we want reliability to be addressed as performance.

Patrick Romano: I meant validity.

(Karen): OK.

Patrick Romano: So, if the – in the case of in any-or-none or all-or-none composite, I think it could be argued that what matters if it's a single concept, as Nancy has pointed out, what matters is the validity of the measure as a whole, not the validity of the individual components that go into the construction of the measure. So, if it's used as a single measure then it simplifies the task potentially, because in each of the components don't have to be yield individually, because there are simply pieces of the whole.

Elizabeth DeLong: This is Liz. I'm trying to follow and I'm sorry I didn't get on until just a few minutes ago.

(Karen): Is that Liz?

Elizabeth DeLong: Yes, I'm sorry. Can you hear me?

(Karen): Yes. I just wanted to make sure who's voice that was. Thank you.

Elizabeth DeLong: But are we discussing whether or not to consider an any-or-none or all-or-none measure composite? Is that what we're doing?

(Karen): Yes, yes. We actually jumped to the last section of the – or almost the last section of the memo about what kinds of measures actually would fit our idea about what's a composite performance measure that would require the two additional criteria that we've been talking about, the one about, you know, under importance of describing the quality construct, rationale, and (inaudible) and storing fit that and then probably more importantly the one under scientific acceptability, the additional analysis to support the measure construction, which would vary depending on kind of the model of those components.

So, we are dissecting, you know – based on the survey that we did, there seem to be two areas of agreement and one with a little less agreement and then the areas that there was either agreement, they were not composites or no agreement, so. But we are talking about the all-or-none or any-or-none types of measures, which are basically computed at the patient level, scored at the patient level, you know, versus taking already scores and combining them and whether those should be in the composite group.

Patrick Romano: And the proposal briefly that I put on the table Liz was that at least the any-or-none composites, which are the most controversial here, perhaps also the all-or-none. But at least the any-or-none should be really viewed as single measures and should be submitted, evaluated, and endorsed just as any single measure would be.

Elizabeth DeLong: Yes. I have trouble buying that and I don't want to be obstructive. But these are individual performance measures that are – that all need to be collected. If you're designing the IP system or the questionnaire or whatever,

each one of these measures has to be collected in order to calculate and compose into one measure (so) all of them. So, it's hard for me to argue that that would be one measure. I mean, it's a composite measure to me, the composite of a bunch of things that had to be collected.

Sherrie Kaplan: Yes. This is Sherrie again. I hate to be (inaudible), but when we were putting together the total illness burden index, we have multiple chronic conditions each of which has a severity score. We put that together as a composite. You wouldn't assume that because somebody has arthritis they necessarily have COPD or anything else.

So, when you evaluate it, the reliability test you have to do is interclass correlation. You have to say it discriminates people with, you know, one's physicians, cadre of a physician's panel from another physician's panel. Do they have sicker patient, so you would do (interclassical). You still have to subject it to the liability testing. It's because it's formative, it's a different kind of reliability testing and anything that's multidimensional items that measure a single construct is a multi-item measure and it is a composite.

And, you know, I still think that the scoring of it, all-or-none, any-or-none, is irrelevant when you're taking multiple items and creating a single measure that measures the construct.

Patrick Romano: Well, I think we may not – we may just not come to agreement on this. I think the example that I gave before you came on Liz was in clinical trials that you and others have worked on, we often have might you might call a composite outcome, which might be based on, for example, having a nonfatal stroke or a nonfatal MI.

And those may be collected through separate methods you might be able to want those two outcomes separately, but when you design your trial, when you power your trial, when you report the trial, you report them as a single outcome. And at the end of the day what matters is the performance (as) single outcome.

So, from the standpoint of NQF committees evaluating a measure, I think that their focus should be on evaluating that as a single outcome measure, not trying to disentangle all the pieces and reconstruct it. As ...

Elizabeth DeLong: I think you always make very, very good point. But I would suggest that when we do these outcomes that consist of several different outcomes, we actually call them composite outcome. And it's rare that we analyze them and don't look at the individual ones in addition.

But I think your first statement was right on target. I think there are few things in this document that we're never going to have total agreement on and this may be one of them.

(Karen): So, I think the question for us, you know, from the survey there seem to be more agreement on this and some of the other change. But, you know, we need to figure out how to go forward. So, do we just focus on the things where we have agreement and have a plan to monitor the other things? Or, you know, if it – there's some substantial agreement, for example, on, you know, all-or-none or any-or-none put them into composite bucket and monitor, you know.

I think we're not going to reach total agreement, as has already been pointed out, but what's that fast forward in terms of, you know, being able to perceive, you know. One of the key issues that we've had in terms of this project to begin with is leaving it up to individual developers, created a lot of inconsistency and perhaps in some cases or in some (eyes) a lack of fairness. And we need to try to do something that we can be consistent across, you know, NQF staff steering committees and developers and, you know, really signal ahead of time what the expectations are.

So, I'll stop this to see what your suggestions are.

Sherrie Kaplan: This is Sherrie again. What are the consequences of a false positive and a false negative for NQF?

(Karen): So, the question of a false positive I think is what Patrick has been laying out that if we don't need that extra information for a particular measure, it's more

information for the developer to provide more information for staff and steering committee to fits through. The false negative means that may be we aren't fully evaluating a measure as we should be. So, you know, I don't know if I know the consequences at, you know. This is not a perfect process or science to begin with, but I guess others may have some other ideas about the consequences.

David Shahian: This is Dave. Nothing that's come out of any NQF taskforce has been perfect and this isn't going to be perfect. I think we need to do the very best we can, try to get consensus. I like the category in the survey you sent around, I can live with it. And I think if somebody really has a strong scientific objection and can make a good case for it that's one thing.

But absent that, I think we ought to do the best we can and come to reasonable consensus and correlate.

Nancy Dunton: This is Nancy. I think the only thing to do in this circumstance is to make the call that will allow developers to know what their tasks are and then monitor the growth of the discussion of the science around this, because if you're going to not have confusion reign in the committees then there need to be calls made about whether this should go through the composite, this (filler) indication goes to the composite process. So, I think clarity and direction is the thing to strive for at this point and just not move forward on the areas or don't include the area in which there's division.

That probably wasn't very clear if I'm calling for clarity.

(Karen): No, I think that's, you know – what we're aiming to do is to, you know, make the best coach, you know, that we can and knowing what we do at this point and minimize any of the undue burden. I think that's the main negative or main consequence that might be a concern is adding more burden when it may not be necessarily.

Female: Then that's certainly a vote in favor of not calling the any-or-none or all-or-none a composite. And imagine most of us can – well, I can live with that if you don't want to call it a composite.

Male: Oh, why don't we see what the majority of members feel about it?

Female: OK. So, we could go down our list on the – so, what we're talking about is another of any – and I don't know if you want to do the separate way or together, the any-or-none or all-or-none, whether those should be considered composite.

And I think we always need to put the caveat that we're talking about this in terms of NQF measure submission, evaluation, and endorsement. So, I don't think anyone would argue that a combination of things whether it requires additional information and evaluation in terms of NQF endorsement at this stage.

So, does that – and I guess are you suggesting that we go at to kind of a roll call and ask for people's opinion right now, David?

David Shahian: Well, I guess I would – I would say that first of all we have to ascertain whether there is an absolute compelling scientific argument one way or the other. And I've heard arguments on both sides of that, I'm not convinced that there is and I personally lean towards calling these composites.

But I like to get a sense of the group as a whole where people are leaning. I think if there's absolute compelling scientific evidence one way or the other then I think, you know, that's one thing. But in this case, I think one can make arguments on both sides that are (inaudible).

(Karen): OK. Given that – I mean, we did ask this question on the survey that I think there has been more discussion and suggestions. So, why don't we just go down our list and see where people do a strap hole, see where people stand on whether this should or should not be considered composites for NQF purposes.

So, I'll just start with top of my list. Patrick?

Patrick Romano: Yes. I think I've stated my views and I believe that this should be evaluated as single measures.

(Karen): OK.

Patrick Romano: They should not be – they should be put to NQF for evaluation, but they should be evaluated as the yes-no concept that they are at the patient level.

(Karen): OK. Liz DeLong?

Elizabeth DeLong: I believe they're composite.

(Karen): OK. John Birkmeyer is not on the call. Dale Bratzler, have you joined us?

OK. Jim Chase?

Jim Chase: Oh, I'm torn by this because I can live with it the way it is. I am still compelled Patrick by your argument about. I could see them being left out but the – I'm also interested in this idea, you know, alpha and beta (errors).

So, I would suggest that we leave them in now, because I think they may reviewed as composites and may be more rapid and it's not as onerous in that review. But I think we should give a little experience with reviewing more of them, so I'm going to vote leave them in.

(Karen): OK. Nancy?

Nancy Dunton: I'm going to vote that both of them are not composite.

(Karen): OK. All right. Liz? Liz Goldstein?

Liz Goldstein: Yes, I would treat them as single measures and not composite.

(Karen): OK. Sherrie?

Sherrie Kaplan: Yes. I vote that they're very definitely are composites.

(Karen): OK. And has Lyn Paget joined us?

OK. Dave Shahian?

David Shahian: Composite.

(Karen): OK. Steve Wright? Did he join us?

Steve Wright: Yes. And both are composites.

(Karen): And Alan Zaslavsky, have you joined us?

OK. So, according to my numbers here, we have one – three nos and one, two, three, four, five yes and several people of course are missing. So, let's go on to some of the other issues, because may be when we talk about the criteria specifically that may shed additional light on this.

But any suggestions that those that's not an overwhelming majority and we are missing a few members' vote there. We can also follow up after this call with another vote on that if that would be useful.

David Shahian: You know, I'm willing to admit that, you know, I can see arguments on both sides and that I – but I can lean towards calling them both composites of the people that tells that they shouldn't be composites. Are there any absolutely rigid empirical evidence that it should be considered as an individual measure or is it more of just that you're sort of winning in that direction?

Male: Well, I think ...

(Karen): OK.

Male: ... (inaudible) our specific implications for some of the guidance as related to the specific criteria.

(Karen): Right. OK. And before we move on, in terms of the list that we currently have that would not be considered composites or purposes of NQF measure submission, evaluation, and endorsement those also are on that same page. And let me see this.

Yes. At least if you move to that list, the following will not be classified; right. Are there any areas that people want to discuss before we move on?

Patrick Romano: Well, I think – I would say I think that you did the right thing to kind of set this aside. The multiple link steps is basically just – as I understand it's basically just all-or-none composites in which the steps are viewed as pieces of the same component or domain of quality. So, that's just kind of a special case of the all-or-none situation.

(Karen): Right. I think that's a good observation and we kind of toy with that and whether, you know, they should be considered all-or-none because, you know, we can look at it that way that may be another argument in terms of how we view this because – but good point I think.

Male: (Inaudible).

Patrick Romano: It's for treating shrinkage or adjustment as a composite. But I got the sense from reading other people's comments and particularly David's comments were helpful that probably that's not the sense of the majority of the committee.

(Karen): Right. And I think, you know, again, that seem to have even more controversy or division or agreement that they were not. So, you know, again, this is not something that will be stepping stone and as we continue to get more experience with those types of measures, we may, you know, in the future think of them differently.

OK. So, I think may be related to the conversation that we're having and where is the most impact. So, going back to the question of what's the consequence of categorizing something as a composite measure, I think, for me, it relates mostly to the additional analysis that we would be interested in for composite performance measures, which comes as item (2d). And Elisa, if you want to move that on the webinar and anyone else that's following on their own documents, it was on Page 4 of the memo – actually it starts on the bottom of page three – sorry.

So, the idea is that, you know, we do have criteria for revival of the invalidity, and you know, we can certainly talk about them again, but the – those are basic criteria applied to any performance measure including these composite performance measures that we've already talked about, you know, we (vastly)

demonstrated for the composite measure as a whole and – but in addition to that, there would be some expectations that there would be empirical analysis to support the composite construction.

So, this is, I think, where the additional burden could be or maybe we'll find that these things just don't fit at all for certain types of measures that we make through this in our prior decision. So, we want from analysis that will show if the composite measures fit the quality construct and add value to the overall composite.

The second one was that the aggregation and weighting rules are consistent with the quality construct and rationale. And the third one is the extent of missing data and how the specified handling of missing data minimizes bias.

So, I think, you know, as we've been discussing these and working on this, those are great idea, but you know, the devil is in the details. So, what does that actually mean in terms of what we would expect. We had some discussion about if the components are correlated and this would be, if we're talking back in terms of model terms, the reflective model.

So, if the components are correlated, then we would have analyses that are based on shared variation, factor analysis, item total correlation (inaudible) in this category, and if the components are not correlated in what could be considered formative, then analyses demonstrating the contribution of each component to the composite score, correlation of the individual component measures to a common outcome measure where some suggestions that were made.

You know, of course, in the example we've been talking about is that if the composite of outcome, such as complications, what would those analyses be. So, I think the question that we're left with is, you know, if these are really criteria, do we need to be able to apply them and give examples of what kinds of analyses we would really be expecting?

So, we always come back to, you know, the reflective of what people are called psychometric model where you can deal with analysis based on correlation to identify whether those components really add value, but how

does – how do these criteria translate when they're talking about the formative model and all or none? Do they really require anything beyond the liability and validity testing of the composite measures?

So, I'm going to stop there and see what you're thoughts are about additional analyses that we would want for composite performance measures.

Female: Well, one thing that I think you may have already covered but that would be the sensitivity of the composite to eliminating some of the components.

(Karen): OK. And so that would be – so for example, we want examples of the number two, in terms of other aggregation and weighting rules. So, you're saying that we could have some sensitivity analyses for the component measures also if it is a (formative) model.

Female: In terms of how they contribute to the overall. If they don't – if leaving one out doesn't change the overall, then it is not necessary to be in their.

(Karen): OK.

Sherrie Kaplan: This is Sherrie. It depends on the purpose of measurement and if they're trying to – for example, the purpose of measurement is to distinguish one physician's quality from another physician's quality or on one – you know, one hospital from another hospital, you should at least ask for some interclass correlation that show you that, you know, there is consistency within the hospital and differentiation between hospital service within the physician and differentiation between physicians for those kinds of formative measures.

Female: Right. But we – isn't that also true for a reflective measure because – I mean that's our basic criteria for reliability of the composite measure. That's what – your recommendation was that for any composite performance measure, they should be able to distinguish performance, and that is the role of any NQF-endorsed performance measure. And I think ...

Sherrie Kaplan: And we can (inaudible) earlier up on the screen, that's exactly right. So, you know, whatever, but that kind of analyses different from the kind of right and

variable analyses for reflective models. That kind of, you know, interclass correlation models often use for these more formative models as well.

(Karen): So, I think I'm going to, you know – I guess I'll push this a little further. If we say that the composite performance measure should have a demonstration of reliability at the overall composite and that we're really focusing on the performance measures for looking at this issue of variability within providers versus between and we say that validity should be demonstrated for the computed composite score that would actually reflect quality, then do we need to ask for additional analyses as we've outlined here or is it only in very special case that it is the (inaudible) only applicable to the reflective model?

I mean in this thing data is an issue – I think a bigger issue for composite performance measure is that it certainly can be an issue for single measures. I think it gets sometimes multiplied when you're talking about composite performance measures. You know, that's maybe a different issue, but I guess, certainly, the one and two ...

Patrick Romano: And (Karen), I'm keying on the fact that we're using the word quality construct in this criteria in (2d). So, just the use of that word implies that there is a latent construct that – and therefore, it really is consistent with reflective measurement model?

(Karen): I think that we can use the term quality construct for formative model as well, can't you?

Sherrie Kaplan: Sure. I mean I would argue that – back to Patrick's point, I think that the issue is whether a composite (add) anything versus an individual measures or whatever the criterion of record is because if you're – if this is about composite measures then even for a formative model, you're trying to – you're trying to add things together that collectively do a better job of measuring quality than some single-item – individual item that either has been used in the past like mortality or broken out helps you in some different way. The composite has to do a better job than something else.

Patrick Romano: A better job, I agree, but is it necessarily a better job at capturing a latent construct? Couldn't just be a better job at maximizing patient utility, helping the patient to balance different competing hazards?

Sherrie Kaplan: I view latent construct as all of what you just said. I mean, you know, quality is the latent construct I think NQF is trying to measure ...

(Karen): Right.

Sherrie Kaplan: ... and only for a specific purpose, there's a different issue, quality for helping patients make decisions, quality for, you know, paying physicians differentially based on performance. Now, all of those things are differential applications, but you know, the underlying construct, I assumed, was quality.

(Karen): Right. I think that's the case. I mean even – Patrick, you still have to come out that, you know, going back to our criterion 1d, what is the quality construct and what's included in that? So, they were about quality to some extent.

Patrick Romano: Right. But the purpose – it's sort of a timing issue, right? So, if we think that there is some unobserved latent construct of quality, but if what we're trying to do is to – is to come in at a later point basically and help consumers with the fact that Provider A has a higher mortality rate, Provider B has a higher complication rate, and Provider C has worse functional outcomes.

And so, our composite may provide a tool to help patients understand how to maximize their utility. Even if after and the quality is separately driving each of those pieces, but the purpose of the composite is not to measure a single constructive quality, it is to help patients enhance utility.

(Karen): But your quality construct could be conceptualized as mostly component that involves function, complications, and whether it survives. So, again, I mean I think it – you know – I guess – yes, we still leave that description of what the composite what the quality construct is; however, the developers is defining that and the rationale and its – how it is going and rating and what's included (next shot) which we do have in the 1d and I think there's a lot of agreeing – there's absolutely a lot of agreeing in on that.

There's actually quite a bit of agreement on this (2d). I am just wanting to kind of see if there's anything else or if you had any other suggestions of, you know, whether we really can expect analyses in these areas of if it is – just in discussion, how does that discussion differ from what we asked for in describing the quality construct rationale and how it is put together.

If people agree that there are empirical analyses that can be done to 1, 2, 3 and what are the examples that we have below that, are truly examples as the kinds of analyses that might be done.

Patrick Romano: Well, certainly, the examples are reasonable. One could certainly think of scenarios where one would want to provide more flexibility to steering committees. For example, in the case of a formative design where you might want to include certain clinical concepts because there's a strong clinical rationale for including them, then, it – some of these analyses wouldn't pertain.

(Karen): So, that would be included in the evidence, start there in terms of why you would want to include the component in a particular composite, and the question is then just because you have that clinical rationale that sufficient then or do we still want them to do some analyses to show that it has additive value.

Patrick Romano: So, the additive value may simply be that it doesn't make clinical sense without it.

Female: I am again sort of confused about the dark side is asking for this. If in asking for this, people provide a very compelling rationale for why they didn't do it, then, you know, you can look at it and say, "Wow, it was either very compelling or it wasn't."

(Karen): Right. Right.

Female: To ask for it, I think it is completely rationale. I mean I'm having a difficult time understanding why you wouldn't want to ask for this.

(Karen): OK. I could live with that.

Female: (Inaudible) and I think, you know, we can certainly – we worked this line that says, “If empirical analyses does not provide adequate result, other justification must be provided.” We can kind of rewrite that so that it is more reflective of this flexibility or that they would, you know, have the provider rationale with the steering committee could either, you know, accept or reject, depending on whether that’s an adequate rationale.

(Karen): OK. Well, I’ll move on. Maybe we can go back to the top then. If you go to Page two, the definition – so basically, there is agreement – there is agreement about the definition and one suggestion for minor change which we made but others may not think it was so minor, but we have two or more component measures, each of which individually reflects the quality of care, but was changed from the entire wording of each of which can be used individually to assess quality of care.

Female: (Inaudible) change.

(Karen): I’m sorry, pardon me.

Nancy Dunton: This is Nancy. I liked the change. I think it is clarifying.

(Karen): Anyone have any objections. OK. Let’s move on to the – go ahead.

Patrick Romano: Sorry, I lost my connection here for a second, but was the change to specify that each individual component had to reflect quality? Could you read it again?

(Karen): Yes. If the component measure is a combination of two or more component measures, each of which individually reflects quality of care into a single performance measure with a single score. The wordings prior to this was each of which individually can be used to assess quality of care. And the suggestion was to instead just say reflects quality of care.

Patrick Romano: OK. Thanks – thanks for the clarification.

(Karen): OK. The next area was defining the quality construct. As you know, what we're asking was for people to describe the quality construct and I know we tend to revert back to the discussion of reflective and formative models or psychometric informative models, but we have decided to not necessarily use those terms but we wanted them to describe the quality construct and we were asked in the commentary to define quality construct.

And so, we did that and had some discussion about that in January and then the revision to the report that this is our latest attempt of doing this and so I'll just let everybody take a look at this either in your own document or on the webinar, you can read all that stuff.

Female: Where are you again?

(Karen): OK. We're on Page Two of the memo, kind of in the middle of the page, quality construct.

Female: Thanks.

(Karen): OK. So, basically, we're saying that the quality – were saying – well, I'll just go ahead and read it, quality of care is an abstract concept that is measured using observed variables. Composite measures are complex, multidimensional, and represent a higher order of construct than the individual measures.

The composite measure quality construct is a hypothetical concept of quality that includes – and this is kind of getting into what we would want them to include in the measures commission, what is the overall area of quality, and we've given some examples here, so for example is the quality of CABG surgery, what are the included component measures, discussion of representatives of the included component measures, what are the conceptual relationships between the component measures and the overall composite, and then – and the relationships amongst the component measures, so are they correlated or not.

So, it's really to be specific about how, you know, doing some description of what the quality concept is. So, the question is, if it's what we'd like to see, if

it's too prescriptive, do you want to just be more general? I think we had actually quite a bit of agreement about this in the survey, but I'll just stop there and see if there are any thoughts.

Patrick Romano: Well, let me try an example to see how this would play out.

(Karen): OK.

Patrick Romano: Because this is just an example of where we were struggling a little bit. So, consider a composite that combines a measure of patient satisfaction and a measure of mortality and let's assume that those measures are completely (inaudible) that the provider behavior is getting increased patient satisfaction are independent of the provider behaviors that lower mortality.

So, there is no single latent construct that's being measured there, but yet a composite might provide a rationale mechanism to allow consumer to trade off competing values between higher satisfaction on the one hand and higher mortality risk on the other hand in a matter that maximizes utility.

So, how do we deal with that issue where there doesn't seem to be a single underlying latent construct and yet there does seem to be some value in the marketplace for such a composite?

(Karen): Well, OK. I'll see if anyone has a response to that.

Jim Chase: This is Jim. When I hear you go through the example I would say – if I were answering that I would say that my construction is it is important for patients to both have a – to have quality of care both a good experience and not to have had any errors or other problems during the procedure.

So, I mean I – to me that fits together on that space. I understand it is not a single construct but it is something that – what we're asking for is the measures (doer) to say why they are putting these components together, what did they – what did they hang together, what makes sense about them, and that one seems reasonable to me.

Male: Hi. My response to this is that would not be a useful composite. The only way that that composite could serve the consumer is if it was deconstructed into its two component domains, and if that's the case, then the composite itself provides no value because you've – your example is one in which they are (inaudible). There's no – you know, they are completely independent and you postulated that both satisfaction and clinical outcomes are important. So I don't – I don't think there's any utility to that composite.

Male: The attempts is mainly construct actually revealed that with the specific components that I think that comment actually points to the value of being pretty open ended about what the construct is, but then looking at it critical – critically, it will be something like overall quality of share or something like that and you just have to argue that you have gotten parts of all overall quality and that you have weighted them together appropriately and probably you have probably you'd fail to make a good argument (inaudible) very good quality measure.

Male: You'll just get a score of 5 by having very good patient communication and bedside manner and very poor clinical outcomes or very poor communication and excellent outcomes, so the composite itself is uninterpretable.

Male: Well, it depends how you're interpreting. In other words, consumers, when they make decisions, they construct those composites in their minds implicitly. So, we are inherently waiting these differences sometimes or (orthogonal) components against each other.

(Karen): Right. Then I guess the question is if you put together composite with your weighting and with the patient can't deconstruct it based on their value set, how would that help them. I think that's one of the questions.

Sherrie Kaplan: This is Sherrie. If you're adding up apples and airplanes, you're in trouble. But if you're trying to say that hospital way is better than hospital (inaudible) because it doesn't kill people and it makes them happy. Then, I think what you're trying to do, Patrick, is make the underlying, you know, conceptual argument that we're asking the measures developers to make.

They better come up with some decent rationale for why they are combining this that's both conceptually and empirically compelling. I mean then they have to really provide us evidence that says, "Look, if you use these things, you get this kind of discrimination between hospital and this kind of consistently within hospital."

Patrick Romano: OK. I can buy that.

(Karen): And we're not saying that the components have to be related, we're just asking, you know, so in that case you would say, "They're independent. They are not correlated." You know, so this is really just for the developers who describe, you know, their quality construct and then, of course, the other components had the rationale which they're getting into and then the – how their – eventually how their construction actually fits the quality construct and rationale.

Patrick Romano: Yes. I think this – I mean I just post it to the stress test because I want to make sure that our framing is broad enough to capture some unusual types of composites that might come to NQF.

(Karen): Right. OK. So, I'll just mention if, you know, after we get off this call, if any suggested, you know, language or clarifications come to mind, please feel free to send them to us and also we can go back, but in the interest of time, we'll just keep moving.

So, again, during the whole process, we talked about – in general, we want to apply the NQF criteria for composite performance measures and just, you know, be judicious and where we ask for additional information, one being the quality construct that we just talked about.

So, in terms of NQF criteria has – NQF has criteria for reliability and validity and the discussion among the top was that we just need to provide some guidance for the purposes of composite performance measures, reliability and validity should really be focused on the overall composite score and that's the distinction from NQF criteria for any measures currently allows assessment of reliability of either the data elements are the performance measure score and the discussion among the text that led to this – these recommendation was that

first of all we talked about we may not have a reliable component that maybe one of the advantages of putting it into a composite.

The second is that even if you have individual components that were reliable and valid, once you put them in your scoring aggregation and rating system, whether you end up with the reliable valid performance measure is still a question that we would like addressed in reliability and validity testing. So, that's my understanding of where we ended up with trying to focus this at the level of composite score and Sherrie was saying about the within versus between variability.

So, if you would go, Elisa, to the top of Page Three, and for those following along in their document, the top of Page Three talks about reliability. And we had some language that was pointed out was too restrictive where we really talked specifically about signal-to-noise analysis that language that seem to be too restrictive.

So, what we're suggesting is the language that testing should demonstrate that measurement error is minimal relative to the quality signal and then give some examples.

So, signal-to-noise analysis that we can put in the references here as described by RAND, interunit reliability has been described by Alan in the paper he did, Sherrie has mentioned the interclass correlation coefficient. So, does that make sense? Is that correct language now or do we need further notification?

Female: It's fine with me.

Male: Yes, I think so as long as the references are provided also to help the developers.

(Karen): Right. We will definitely do that. If you have some other suggestions, you know, please run this now. And then at the end, we had talked about reliability of the individual component measures is not sufficient and in some cases component measures that are not independently reliable can contribute to reliability as a composite measure.

This was kind of our rationale for two things. One is that we don't need to require reliability of each of the component included in the composite and then the rationale for look at it at a performance level.

OK. And then for validity, I think we had some areas of controversy. I think the basic concept was the same, but again, once you've go through all these aggregation and weighting, do you end up with a valid representation of quality and that should be done for the composite performance measure.

On the last call, there were some discussions about whether that was realistic to expect at the time of initial endorsement. And so the decision was made that we should accept validity as the component measures but must meet into our criteria and by endorsement maintenance to have some of its validity testing of the composite performance measure demonstrated. So, it gave them a little more flexibility.

So, there are some concepts our – one comment whether that was even realistic to expect validity testing by the time of endorsement maintenance because it is unlikely that a gold standard exist. So, I think that's true. It is unlikely that a gold standard exist.

So, generally, validity testing would focus on constant validation, not expecting that there will be some other criteria in the measure of the same quality construct that you would be able to do validity testing with.

So, I guess the question is whether, you know, we're still in agreement that we should ask for validity testing as a performance measure score, whether that should be an option that it's the individual components or the composite score, or whether it is just totally unrealistic expect validity testing.

Male: It definitely depends on what the basis of validity is. You know, if you have a performance measure whose fundamental construct is (inaudible) that contribute to a better outcome and we might initially endorse something on the basis of a literature that says that's two or three significant process and maybe ask for – announce this as more data is collected that shows for these as variables or something like that. But if validity is based on – this is

(obviously) tested then it is not clear if there's any quantitative thing that would add to that.

Female: For me, it depends where we are in the course of the development of the new measure than the composite because if you're at the very (outset), as Alan said, face and content validity is maybe what you're stuck with.

But on the other hand, if there's been some experience in the literature with either the individual components or some collective of the components, then it is reasonable to expect some empirical testing evident. I would hope that you'd set the bar pretty high and at least ask for a rationale for not doing it.

(Karen): OK. All right. Any other suggestions or thoughts about this? For those of you – so I think one question that comes to my mind with this is, for example, some of the art composite measures that are basically combinations as outcome measures, what kind of construct validation could be envisioned in your composite is that composite of outcomes versus processes? What kind of theoretical concepts or hypothesis testing would actually be relevant to that kind of composite performance measure or is that one of the cases where you might rely on the patient content validity?

Female: I hate to be the nosy one again, but to me, if you're going to create, for example, a composite of lipid levels, hemoglobin A1c levels, and blood pressure levels, and call that quality of diabetes care, then I think you have to subject – and for the purpose of discriminating one clinic from another or one doctor from another, then you'd have to provide some empirical evidence that that collective has some value in doing that measurement task and at least it is giving you more evidence of diabetes quality than each of those indicators taken separately and there is more, for example, interclass correlation when you take them as a group than when you do them individually.

Male: So, I guess there was – what would they have – what would the major developer have to submit then to prove validity in your mind? I mean because it is not going to be feasible to look at, you know, outcome measures for the patient in a short timeline. I supposed you could look at published studies, but I'm not sure that really represents validity testing.

Female: Well, the issue of are you measuring what you think you're measuring with those that of collective versus the individual measures, if you think you're measuring diabetes quality, what do – you know, what do doctors who have good diabetes quality also has besides those outcome measures.

Well, you could argue, they do more stuff. Now, that's not independent because you're measuring your annual hemoglobin A1c. You can't have an outcome unless you had the process. So, that doesn't get you cleanly out of that situation. But actually, we have that same issue when we published the article on using NCQA data and the DPRP on discriminating individual physicians.

So, you could argue that, "Well, I'm going to take these as a collective and I'm going to see to what extent that collective is related to other things that doctors who deliver good diabetes quality should do.

Female: Right. OK.

Female: Hi.

Female: Go ahead.

Female: I was just going to beat the drum that everybody has heard me beat before which is I think ultimately we need data that demonstrate that what ever has been constructed as a performance measure measures performance in such a way that it is related to outcome and that it is consistent across time which requires demonstration that there are data to support that.

Perspectively, we've decided not to do that, but once a measure is actually approved, I think we need data.

(Karen): OK. Just a question about the consistency across time, one of the – actually, this is a little off topic but it's something that comes up, sometimes we have measures submissions where for reliability was not to demonstrate consistent across time and the question comes up with in the context measures being put forward for performance and performance improvement is that a reasonable

expectation that it should be consistent across time and how do you factor in that the goal is change over time?

Female: No, I absolutely agree, but there is some relative ordering that you would expect to maintain itself over time. You're not going to expect the worst performer to become the best and you're not – if it's not demonstrating a relationship with outcome, then outcomes are going – should scale over time as well.

(Karen): OK.

Male: I guess where I have trouble with the last discussion though about outcome is that I can see – but perhaps using that, if I was doing a re-endorsement evaluation after the three-year period on a measure, you would expect that perhaps there might be some outcome data available. Where I struggle with is in the initial submission.

So, somebody – one of you used the diabetes care as an example linking two or three different performance metrics around diabetes that have been associated with improved patient outcomes and perspective trials. But when the three majors are together, I actually like the concept of saying, “Well, you know, if somebody is doing well on this measure, there is also probably somebody that's doing well on other measures.”

But actually, being able to demonstrate in an initial submission, that there are outcomes data for this composite that it reflects – it actually is the link to better patient outcome, I just think that's a high bar that allows that to be able to make.

Female: I think I gave in on that one last time.

Male: Yes.

(Karen): Right. So, I think what we said in the last time was that – yes, that was probably two nights – maybe – depending on the measure and the circumstances maybe too harsh to expect the initial submission and that the component measures should be demonstrating as needing NQF criteria for

validity but what I kind of heard today and maybe is more appropriate is that they would be able to share contents or face validity of the composite performance measure. And so, would it be either way that if they had validity of the component measure that would be acceptable or content or face validity of the composite.

OK. Well, I think we'll try to be flexible here as we put together this final so that I think that we're testing some of the questions that we're hearing to, you know, indicate that would be good measurement science that to be flexible in realities of the real world.

OK. If we move down on Page Three, this was the additional information on composite performance measures that we would ask for – we've already talked about the quality concept. The second one was the rationale for constructing a composite measure and then the third one is how the aggregation leading to the composite – of the component measure are consistent with and representative of the state of the quality constructing rationale.

So, this would basically be, you know, the measure developers' description of these – in these three areas and the steering committee would really be looking at this for clarity and logic that makes sense to them. Any comments or questions or suggestions about this? There seems to be lot of agreements on the survey in this area.

OK. And we talked about (2d) already. I think the last thing in the memo and I apologize but (inaudible) ...

Patrick Romano: (Inaudible).

(Karen): Go ahead.

Patrick Romano: Yes. I was just going to say you're looking at the bottom of Page Three, right? Or near the bottom of Page Three?

(Karen): Yes. Yes.

Patrick Romano: Yes. The only thing I would take in that because people often interpret a construct as a single latent construct in the psychometric model, so we may want to explain that we're using the term construct in a somewhat broader way.

(Karen): OK.

Patrick Romano: I'm not sure exactly how to do that. I could try to write a sentence or two.

(Karen): OK. So, Patrick, this would go back to and certainly if you have some suggestions for this, but if we go back to Page Two – Elisa, if we could go back to the definition of the quality construct and you may want to send – this would need to accompany it but where we say it is an abstract construct that is measured using (inaudible) variables. Composite measures are complex, multidimensional.

So, we are kind of thinking that the quality construct has multidimensional components. So, I don't know if you have some suggestions on the way got this wordings currently or – but we would welcome any clarifications to make sure that people are all on the same page.

Patrick Romano: OK. I'll try to suggest tweaks.

(Karen): OK. Great. And then if we go to the last area that we had quite a bit of discussion about and is about decision making context that's part of the rationale that's in the report, and this is on Page Five of the memo. So, what we have gone back and forth about this and I think this is another area where there won't be complete agreement.

And so, we tried to basically make sure that all views were represented. So, we basically kept the language that was in your last draft. We put the examples in a footnote just to – so that it would flow quicker and would be more concise and then we – the last paragraph tried to make sure that all of the viewpoints that were expresses by the text were included.

So, the language that we had previously, it is still in there but some TEP members thought decision making context is a unique aspect of composite

performance measures with choices about which components are included and how they are aggregated and weighted can affect interpretation and use.

And other members expressed concern that it is inconsistent with NQF current policy to endorse measures feasible for performance improvement and any accountability application or that another of you point that it is unnecessary because all composites should be a valid reflection of quality.

So, I will ask you to kind of think about that whether that language makes sense that it accurately reflects the various viewpoints that have been expressed about decision making context. I think that the practical implications of this is that we will be asking the measure developers to give us their rationale for their measure which – and they are weighting and aggregation scheme which could include something about the decision context it doesn't have to and it doesn't change the fact that NQF currently endorses performance measures for any accountability application – not specific accountability applications, though I will say that is something that, you know, there's a question that come up in general about NQF endorsed measures and something that maybe looked at in the future of a different way to go about that but I think we just want to make sure that we adequately expressed the various view points on this.

So, I'll stop there and see if anyone wants to add anything or you know definitely you can send it to us afterwards as well. So while we're having a chance to digest that, why don't we see if there's any comments or questions from the audience, from the public members, public and NQF members on the call, that maybe on the call.

Operator: At this time if you would like to ask a question or have a comment, please press star one on your telephone keypad. We'll pause for just a moment to compile the Q&A roster. Again, that was star one for questions or comments

(Karen): OK, so I'll ask the – we'll talk about next steps but before we do that, I wanted to see if there were any other issues or questions that TEP wanted to bring up based on looking at the survey of the TEP members or the – the memo or anything in the – in the draft report as it currently stands. OK.

Operator: There are no public comments or questions.

(Karen): Did you say there is one?

Operator: No ma'am, there is none.

(Karen): OK, thanks, and also if our discussion today – if there is anyone wants to revisit the discussion about whether all or none, or any or none measures should be a third composite at this point in time. So, I will stop there and see if there is any additional items from the set.

Male: So (Karen)?

(Karen): Yes?

Male: If we go back to the material on the bottom of page five.

(Karen): OK.

Male: So, just help us understand what the implications of this. I think the reason that it came up in my mind is – is not that NQF would be endorsing a composite only for certain accountability applications and not for others but simply that the waiting of the components of the composite might differ according to the intended application. And so in the process of steering committee review, it might be important for the steering committee to understand what the developers intent was as far as how the indicator should be used to inform decision making because often the waiting scheme will follow from that.

(Karen): Right, right, so I think that the only implication is this is this the language that will be in the report about our definition about decision making. How this translate into the criteria is that, when we ask for the developers rationale, that's where they would have the opportunity to discuss this and then also the item about how their score – their aggregation and waiting is consistent with that.

So, it really does not have any implications beyond that. At one point, we talked about whether the (inaudible) wanted to actually make the

recommendation that NQF should consider endorsing measures for specific applications, then I think that didn't get a lot of traction and so, you know, and again as I said, that's something that NQF is going to be addressing anyway but I think where I am coming from is basically that this is the language that we would put in the report as – it's just part of the discussion in the – in the narrative discussion of the report.

Male: OK and Alan you raised some interesting points in e-mail and so perhaps there might be a minute to discuss your ...

(Karen): Sure.

Male: ... suggestion now.

Alan Zaslavsky: Yes, I – you know I realized as I commented to Patrick, you know, we joined with an organization that operates under a certain constraints, so I – I generally expect that this committee would be able to change the way NQF does everything but, you know, my point there was that, really what we're talking about is not a single thing when we talked about composites. We're talking about a – a number of different types of things and the implications for additional criteria and the kind of expertise that are needed that are somewhat different, depending on which of these features are present in a given measure.

So another approach to the whole thing, you know, which I have given a spirit of more concept than of trying to change way NQF does seem right now, would – would have been to have a series of check boxes that, you know, not – is this a composite measure? Yes/no but does this measure involved combining several previously existing measures, does this measure combined data from distant sources?

Does, you know – a few different things like that which taken together defined a number of different versions of what it means to be a composite and – and then the – those could trigger the appropriate descriptions of what's required for validation of the measure or, you know, for evidence of the measure at the start or for review of the measure, you know, after it's been in placed for a while which – which might give you a way of having some more specific

guidance than you get just by saying it's a composite and trying to find then criteria that applied to all kinds of composites.

So that's – that's basically what my commitment was about. You know, I was really worried that we could be caught in a logjam over trying to come up with the a general purpose definition of what is and what isn't to composite. Although, we seem to have made it through the call OK, so. I'm glad – I'm glad to know that – that's not an issue.

Male: I guess not – of linking that was that – so there maybe if we think of those two bullet points that – we may still have under the examples of composites.

(Karen): Right.

Male: Would the specific criteria and sub-criteria, it would be triggered potentially differ for those bullets. In which case, I think that we could kind of set aside the semantic argument and not worry about whether all or nones are composites or not if we agree that – that all or none do have to provide some specific information, it's appropriate for all or nones but perhaps not all of the sub-criteria that are specified for other types of composites but I think people are exhausted.

(Karen): Well let's give that a little more thought. I mean we kind of talked about that as we were discussing that earlier in that – relating it back to the (2b). Those criteria about the – I mean I would think you would still want the description, right?

That we talked about in terms of quality construct, the rationale, and the – how the aggravation and waiting tipped that but scenes like that would be appropriate for the second bullet as well as the first but it seems that the question that you heard earlier was more about the additional analyses that we talked about under (2b), whether those are appropriate or the all or none are ending them with that.

Male: So, you know we've been – we've been so busy trying to see whether or all nones are composites but we – I don't know that we talked a lot about what the criteria would be for all or none but they – they might for example be a

criterion that you would only use an all or none for a set of closely related processes that are part of the same overall, you know, shared objectives, the same overall objectives or something like that.

(Inaudible) doing overall amount of a bunch things that are really would be coupled that are done by different actors for different purposes and view that as a quality composite. Although, you might be wanting to use the overall rate of good care with the quality composite. So you know, there could be in both less criteria and more criteria for all or none. We haven't really hatch out when we think all or nones are to be considered appropriate. So, it's a little hard to say that what those would be.

Female: Can we visit this is one of those don't try this at home kind of phenomena where, you know, if you don't head your screwed on right and provided a compelling rationale for what you're putting together and why you're putting it together, the purpose you're putting it to.

You're ending up, you know, in a – in a complete blind soup and I think it's a whole separate call to figure out under what circumstances and all with non-scoring of a set of who knows what variables might be appropriate and which kinds of circumstances, I think we're now getting into a weeds and until we got more experience with what comes in to NQF, I think – I think, Alan, we might – we might be way ahead of ourselves.

Male: Well, we know a lot about what comes in to NQF. You know they ...

Female: Not on composites, we don't.

(Karen): And we haven't had tremendous amount of experience with them and I think the – maybe the idea – I mean they would still need to – I mean the way we have it now is not to be prescriptive of saying when you can or cannot consider an all or none, that would be up to them making their case by describing that quality construct and their reasoning and then actual – see how the other criteria, you know, do they have evidence or the components. You know, how did those components add value?

So, you know, I – you know maybe as we get more and more experience, we might be able to say the all or none is only appropriate in a certain situation. I don't know.

Male: Well, you could say that someone presents you in all or none should present an argument for why that's a useful scoring in either for you know consumer purposes or – or ...

(Karen): Right.

Male: Or, you know – it's get a little more specific about what has to be justified there.

(Karen): OK, all right.

Male: So then you – you're noting that when it comes to that type of scoring in addition to all the sort of generalities that we have about study why things go together in a composite that the all or none scoring should have a particular rationale and then if there were examples, the examples would be things like instead of (inaudible) which has to be accomplished in order for the overall objective to be discussed (inaudible) and things like never allowable types of events. Those are some of the rationales for an all or none scoring.

(Karen): OK.

Male: But as Patrick said, we are a little late in the – in the game to pick up that kind of thing. It's slightly – you know it could be the same thing could be done in – it's just been in the example and explanatory text.

Male: OK.

(Karen): OK.

Male: Could you (inaudible) to be more directive though.

Karen Pace: OK. So, this is Karen Pace again. Where we're at with this is that we need to pretty quickly finalized this and get it to our (CSAT) who will be reviewing it at their March 20th meeting, in-person meeting and we will need to find out

the agenda time, so that we can have Patrick and Liz available hopefully and certainly any of the TEP that (we) should dial in for it.

So, I think what – what we will do is review the discussion that occurred on this call and indicate and get back to you anything that we think indicates a change that needs to be made in that clean version of the report that we also sent you and we will get that to you pretty shortly next week some time and so we need – if there's anything that you want us to consider or language or suggested changes or addition or clarifications, we would need those from you very quickly, hopefully by the beginning of next week.

So, any other questions or comments, or things that we need to be sure to follow up with you on?

Patrick or Liz, any final words of wisdom for us as we're trying to finalize this?

Patrick Romano: Well, I just say this has been, you know, a challenging discussion and I think, you know, I really tremendously appreciate what everyone on the committee has contributed because we certainly have a robust discussion. I think we've learned from each other. We may not agree completely at the end of the day but I – I think we've reached agreement around certain core elements that hopefully will be helpful to NQF going forward.

Elizabeth Goldstein: And I second that and I think Patrick was right when he said we're all exhausted but you manage to keep us until and you seem to be to have a lot of energy for this. Thanks.

(Karen): Yes. We really appreciate it. It has been very good discussion, raising lots of issues that we definitely need to consider and as David Shahian mentioned earlier, as I think others have seconded, you know, we know that this is going to be an evolution. It's not going to be perfect now.

It wasn't perfect, the first one, and we're hoping we're making progress but we know that will need to be revisited again and we appreciate all of your contribution because I know – you know, it's been time consuming, short timelines and I really appreciate all of your input and suggestions.

And you'll be hearing from us yet again, so we'll thank you in advance for taking a look at what we send out as the final.

(Helen): And this is (Helen), I've been looking and I just want to say my thanks as well, what a great discussion and special thanks to the Karens as well, so.

Patrick Romano: I would – I would say to my colleagues also that please submit specifics, suggestions to Karen and (Karen), if you have ideas or brainstorms that have come up, particularly Alan's idea, it would be helpful I think to see them on paper.

(Karen): Great. OK, well thank you all and have a good evening.

Female: Bye.

Female: Thank you, bye.

Patrick Romano: Bye.

END