UNITED STATES DEPARTMENT OF EDUCATION

+ + + + +

RACE TO THE TOP ASSESSMENT COMPETITION

+ + + + +

PUBLIC AND EXPERT INPUT MEETINGS

+ + + + +

Thursday, November 12, 2009

+ + + + +

The meeting commenced in the Mystic B Conference Room of the Embassy Suites Hotel, 207 Porter Street, Boston, Massachusetts, at 10:00 a.m., Joanne Weiss, Director, presiding.

EXPERTS PRESENT:

HENRY BRAUN
GARY COOK
LAURESS WISE
JIM DUECK
JEFF NELLHAUS
SCOTT MARION

STAFF PRESENT:

JOANNE WEISS
ANN WHALES
JACQUELINE JONES
JUDY WURTZEL

CONTENTS

CONTENTS (CONTINUED)

CONTENTS (CONTINUED)

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

10:01 a.m.

MS. WEISS: (presiding) Good morning, everybody. Thank you so much for coming. We really appreciate it, and thank you to Massachusetts for getting those of us from D.C. out of our hurricane. We appreciate it.

I want to start by giving you a quick overview and setting the stage a little bit for what we are going to be doing today. But before I do that, let me ask -- our plan is to just sort of have a roundtable discussion. Now I am worried that those of you whose necks are craning back there have a hard time seeing us.

The podium is kind of blocking them from being able to see. Well, we will deal with the podium because we don't actually need it until the end of the day.

Oh, perfect, Anya. That's Anya. I'm going to introduce her in a minute, but she's clearly introducing herself right now.

Great. Thank you so much. I think that will help open the room up a little bit.

So, again, thank you all for coming.

My name is Joanne Weiss. I'm at the Department of Education. I am joined here by my colleagues Ann Whalen, Jacqueline Jones, and Judy Wurtzel.

In a minute, we will introduce the experts that we have invited today. But I thought I would start by just giving you a couple of minutes of highlight and framing for what we are doing today.

The first thing that we wanted to do is that we wanted to make sure that it was clear the distinction between the Race to the Top competition and the Race to the Top assessment competition. Hopefully, you all know that you're here to talk about the second and not the first.

(Laughter.)

The Race to the Top competition, we did announce the final regulations for that

this morning. So they are available up on the Department's website today. They won't be in The Federal Register until next week because of the nature of publishing things in The Federal Register, but they are available on our website today.

That is a $4 billion competitive grant that is designed to encourage and reward states that are implementing comprehensive statewide reform efforts.

We did set aside about $350 million to support consortia of states who are implementing common standards by helping to fund the development of the assessments that would be needed in order to make those standards a reality in classrooms.

The applicants for this particular grant are consortia of states. It is noteworthy because you will probably hear us talking about it a little bit up here today, that 50 percent of the funding by law has to flow through to districts in this development grant. So one of the things we are going to

be talking about is, what does that look like and what is the role that districts might play in this? Because we can direct those funds and direct the use of those funds at the district level. So those are some of the questions we'll be talking about.

The process works for this competition differently than it is working for the Race to the Top state competition. In that competition, we at the Department put out proposed guidelines. We then put them out for public comment. We received comments. We synthesized, analyzed, figured out what we were going to respond to, and put out a final set of notices. That is what is coming out today.

The assessment competition is working differently. In this competition, we felt that we at the Department didn't have all the knowledge and expertise. Whoever does? But in this case, we really didn't have the knowledge and expertise we needed to develop the right kind of notice.

So we, with the help of our colleagues at the Department, have developed a different kind of process for this. That process is what you are seeing played out here today. This is the first of several meetings that we are going to be holding around the country over the next few weeks to get expert input and public input at the front end of the process, instead of the back end.

So that when we come out with a notice, it will be well-informed by some of the best experts in the country, but done in a fashion and in a forum that is public, and where everybody can see and hear the conversations that we are having.

I will talk a little bit more about what we are publishing and posting, and where, and how in a second. But that is sort of the gist of how this process is going to be working.

So now let me talk about a couple of the goals we have for the assessment program. To tee-up these conversations, we

did put out a notice announcing these expert input meetings. In that notice, we said, "Here are some of the things that we think are knowns and givens or constraints, and here are all the questions that we have."

What I'm going to talk to you for a minute about now are just sort of highlights of those knowns and givens. You are welcome to go up on the web and find this notice, if you haven't already, and you will see the whole list of them. But the highlights of that are these:

First of all, what we are trying to do is really support states in delivering a system of more effective and useful assessments than we may have been using to date. Those assessments have to provide more accurate information about the students' achievement against the common standards that are being assessed, individual student growth, and whether individual students are on track to be in college and career-ready by the time they graduate from high school.

Those of you who know assessment know that some of the reasons that we've called this panel is that some of these are actually at odds with one another; some of these goals are not the kinds of things that one assessment can produce both types of information. So we are struggling with a number of questions about how one might design a system of assessments that gets us the information that we are looking for.

We also want to make sure that whatever the assessments that we create at the end actually reflect and support good instructional practice, and we want to make sure that at the front end of the process we are thinking through the implications for designing assessments that are accessible to and measure the achievements of individuals with disabilities and English language learners. So some of our meetings are specifically targeted around those questions.

In all of our meetings, however, we have experts who are playing sort of double-

duty who are both handling general assessment questions, but also with a specific lens on students with disabilities or English language learners. So you will hear people around this table talk about those issues, even though we have meetings in Atlanta and Denver coming up that specifically target those questions.

We also want the information that comes out of these assessments to be able to inform both improvement of teaching and learning and programs, but also determinations of school effectiveness, determinations of principal and teacher effectiveness, and determinations of individual student college- and career-readiness.

One of the things that we admit is sort of uncomfortable in the timing here is, and you will see this a little bit reflected on the next slide as well, the Elementary and Secondary Education Act is up for reauthorization, but Congress hasn't yet taken up this question. So we are operating within the current ESEA framework on the things that

we are doing here, but want to make sure that whatever the tests we develop are will be able to be used and cross the transition as we move to the next ESEA. So we need to make sure that these tests are producing the kinds of information that we know ultimately we will need to support whatever directions we are going with that.

So that leads us to the next set of requirements, which is that, at a minimum, we need to still support what ESEA today requires. But we can go beyond it, if there's things beyond this that we think we need to do.

So that means that what you will be hearing us talk about is that, at a minimum, we need to support tests that are in reading, language arts, and mathematics that are annually in grades three through eight and high school, and at least once in high school.

We are developing summative assessments, but that does not mean necessarily assessments that are given on a

particular date once at the end of the school year. They may be given at different times in the school year. They may be given more than once throughout the school year. There may be multiple tests that are given at different times, different grades, whatever.

So those are some of the questions that we have asked for our experts to help talk to us about today.

The other thing is that we are thinking of tests that could replace rather than add to the assessments currently in use. So we are not talking about an additive system. We are talking about a reconceptualization. Of course, it is critical that whatever we do, because these could be used ultimately for accountability purposes possibly, is that the tests have to be valid, reliable, and fair. So you will hear talk of psychometric issues that we have to wrestle with with all of this, as we are thinking through what we are doing here.

So, because of that, we have three

really big-picture goals for the meetings that we are running over the course of the next few weeks.

The first is that we are really trying to paint a vision of what the next generation of assessment systems could and should look like for the country. One of the things that I think is a problem is that our vision today is very rooted in multiple choice bubble tests, and we don't even, as a country, know what is it that assessments could do, and if they look different, what would that mean? Why should they look different? What would that mean for teachers? What would that mean for students and parents? What kind of information could you provide? And how might it change instructional practice in classrooms, if the assessments look different?

So painting sort of the big-picture vision, which, admittedly, whatever notices we come out with next March we are not going to be good at doing because our notice format is not a good vehicle for communicating this. So

we are hoping that through these meetings a vision that's different from the current vision starts emerging.

Second, though, we also want to be very concrete. So we at the Department have a task to do as soon as these meetings are over, which is put pen to paper and actually put down on paper the requirements for a competition that will result in the creation of the kinds of assessments that emerge throughout these conversations. So we are going to ask our experts to be very, very concrete. Like, if you were us, what would you write down? So, hopefully, we'll hear some really specific recommendations and guidance from them.

And the third thing is that we have with us today, and will at a number of the meetings across the country, a number of states. We have done specific outreach to all of the states and invited them to come and participate, because, in the end, the states are the ones who have to develop the proposals

in response to the notices that we have put out, and we certainly want them to get the benefit of all of this wisdom that we will be hearing over the next few weeks, so that it can help inform their proposals.

So, with that, let me talk to you for a minute about the agenda for the day. We have got six experts up here with us. Each of them is going to have 20 minutes to answer the questions that we have put forward in our notice and then 10 minutes of clarifying questions that we up here will be asking of that person, to make sure that we really understand what they are proposing. So it is about 30 minutes per expert. That takes us through the morning.

We have lunch for an hour from 12:15 to 1:15, and I think, for those of you who are interested, they are setting up buffets and things out here for you.

We have two more experts who will then speak when we come back from lunch. After that, we are going to have kind of a

roundtable, almost fishbowl-like conversation about what we heard because, again, our experience is that presentations only take you so far; you really need to have a deeper conversation to really understand what good solutions look like, and an interchange of ideas, experts with one another, and us asking the kinds of questions that we need to ask in order to inform the work we are doing.

We are -- and I'll get into this in a minute -- going to also take questions that you guys have. So I believe that all of you were given notecards when you registered. If you've got questions that you think are important ones throughout the day that you would like us to ask, write them on the notecards. Certainly, we will collect them from you at lunch, but anytime you have an emergency question, you can fill free to run it out to the registration table, and they will run it into to us. We will work those questions into the conversation to the extent that we are able to, and that the questions

are consistent with the direction that the conversation is going at that point in time.

We will then have a quick break, and we have an hour for public speakers. We will move the podium back in at that point in time. People who registered ahead of time as public speakers I believe got numbers this morning and know what order you're speaking in, and we will recap this at the break, so that you know exactly what you need to do to line up and get ready for the public input part of our meeting.

And that's it for the day.

A couple of housekeeping issues then. We just talked about how to submit your questions.

We do have all kinds of timekeeping things going on up here. So, if you occasionally see flashing lights and hear beeps, it is us telling the experts that they need to start wrapping up their presentations. Those of you who are going to be doing the public speaking will also see

little green, yellow, and red lights flashing on your podium, so you will know when you are running out of time.

You will have to forgive me if I am a little bit of a brutal timekeeper. It is really important for us that we have the time to hear from each other and from all the people who signed up to speak to us. So we will be trying our best to hold to those times.

We don't have one of those cute little movie theater ads about putting your cell phones on vibrate, but please do, if you don't mind.

A couple of other things. The sessions are being transcribed today. They will be posted on our website as soon as we can. In addition, all of the presentations that the experts are making will be posted shortly on our website. So you can check there.

We are also accepting written input from anybody, experts, members of the public,

anybody. This is the address that you should submit your written input to, and that also will be posted on our website.

I just want to assure everything that our teams are reading all the materials that you send in to us. So please do submit things, if you think there's things that are important for us to know, understand, or consider as we are writing this notice.

So, with that, I just wanted to end for a minute with a recognition and thanks to all of the people from different states who traveled here for these meetings. They are sitting at the tables at the front of the room. I wanted to just thank them. These are the different states that have RSVPed as having people who are going to be here with us today.

And this is the list of the experts that we have invited to speak with all of us today. I could go into extensive biographies on all of them because they are very distinguished experts in this field. We have,

instead of taking our time up with that, decided to put their biographies in the programs that you all received. They can do a quick job of introducing themselves, if they would like, but our real goal is to hear from them, not about them.

So, with that, I am going to turn it over to Jeff Nellhaus as our first speaker.

MR. NELLHAUS: Good morning, everyone.

I'm Jeff Nellhaus. I'm the Deputy Commissioner of the Department of Elementary and Secondary Education here in Massachusetts.

Being from Massachusetts, actually, having been born in Boston and living in Boston now, I just want to welcome you all here to our great city of Boston, Massachusetts.

You will probably pick up the fact that I was born here by virtue of the fact that I have a Boston accent. It comes through every once in a while.

But I just want to say, first of

all, Joanne, thank you for inviting me. It is a real honor to be able to share with you some of our ideas today.

As you know, in Massachusetts, we have a program called MCAS. I think we would all agree that it has received a modicum of success over the past decade or so of administering that program. But we truly believe that we can make some improvements to that program and really welcome the opportunity to do so, and welcome the opportunity to share some of our ideas of how we could do that with you today.

Also, before I begin, I would just like to recognize Kit Viator in the audience here. Kit is our Director of the Student Assessment Program here in Massachusetts, and the Assistant Director, Mark Johnson, is also here. Kit assisted me in developing this presentation. I would just like to thank her for that and for all the good work that she does on our Student Assessment Program.

Okay. With that, let's just get

right into the presentation.

I would begin by saying it goes without saying that student assessment systems are a critical lever for helping students to success and, in particular, helping them graduate from high school college- and career-ready. Assessments do this by serving two overarching purposes: ensuring accountability and improving teaching and learning.

It follows, then, that the next generation of assessment systems must address both of these purposes. This slide identifies some specific uses of assessments within each of those purposes.

For ensuring accountability, we have such high-stakes uses as determining school and district effectiveness, certifying students for high school graduation, and potentially contributing to determinations of principal and teacher effectiveness.

Under improving teaching and learning, we have such uses as providing models and information for continued program

improvement, differentiating instruction for individual students, and adapting instruction in real time.

Achieving these goals requires a unified system of summative, benchmark, and formative assessments based on common standards, not a single test.

The problem is that no single assessment is suited to serve all of these purposes well. Therefore, what is needed is a unified system of assessments, benchmark, formative, and summative assessments, each designed for a particular set of uses.

In this chart, I depict and contend the best ways to use summative, benchmark, and formative assessments. Just to orient you to the slide, the green dots indicate where a particular assessment is best suited. The yellow dots indicate where a particular assessment might be a contributing source of information, but not the primary source. And the red dots indicate where a particular assessment is not well-suited for the

particular use.

Looking down each column, the chart shows that summative assessments are best suited for accountability purposes, while benchmark and formative assessments are best suited for improving teaching and learning.

Of the three types of assessments, only formative assessment can really assist in adapting instruction in real time. While benchmark and formative assessments could be used for accountability, even local accountability, if you will, it is my contention that doing so would undermine the usefulness. I can answer questions about that later.

Benchmark assessments may be a contributing factor in determining effectiveness of teachers and principals, but only in terms of how well teachers are using those assessments as a tool, and not how well students are performing on them. So the main point here is that summative assessments should focus on accountability. They can

contribute, certainly, to improving teaching and learning. Benchmark and formative assessments are primarily used for improving teaching and learning, and not for accountability purposes.

Now this slide proposes a way in which a unified system of assessments could be designed where the summative assessment serves as an anchor to derive the quality, integrity, and transparency of the system.

The key to this design is the release of 50 percent or more of the items from the summative assessment after each administration. While this will certainly add cost to the ongoing maintenance of the system, the cost will be more than offset by the following benefits:

Item-level reporting of summative assessment results will help increase utility of summative assessments for improving teaching and learning.

Similarly, the release of rubrics and benchmark student work associated with the

release of constructed response items will also promote improvements in teaching and learning.

Most importantly, the released items would be used to generate a high-quality, robust item pool for constructing benchmark and formative assessments. As a result, teachers and school leaders would have the ability to interpret the results of the benchmark and formative assessments in relation to the summative assessments, and I will give an example of that.

So I think the issue around benchmark and formative assessment is, what do we make of the results of those particular assessments? So, if we actually use items generated by the summative assessment, a teacher could answer the question like this: what percentage of the items on the benchmark assessment did the student answer correctly that students scoring at the proficient level on the summative assessment tended to score correctly as well? So just to give you an

idea of how we can interpret the results if we embed in our benchmark and formative assessments items from the summative assessment.

Other benefits include increasing the integrity of the summative assessments used for high-stakes purposes. Certainly, I think we will all agree there may be some security risks of using the same tests or even the same forms year after year, and there is going to be a need to release test items.

And finally, to promote the transparency of the system, as the release of items provides educators, students, and parents a more concrete way to understand performance expectations.

So this is the main feature of the system that I would propose. Okay. So we have some additional considerations beyond the system design that I just put forward.

So, in addition to being a unified system of assessments, this slide identifies several other critical considerations for the

next generation of student assessments.

First, we need to move the mode of assessment administration online. A number of states have already made significant strides in this direction, and more will do so in the future, as a growing number of schools get access to broadband and to relatively inexpensive computers.

The advantages of online administration are obvious: faster turnaround time for results; reduced cost for printing, shipping, and receiving materials, and the potential to develop innovative items, not to mention that doing things online is increasing how students learn and demonstrate their knowledge and skills.

Secondly, the assessments must address high-order skills and not just the fundamentals. We have a lot of experience using constructed response items for this purpose and understand their utility. We have less experience, however, using what I would call performance tasks and tests that would be

administered not in the on-demand environment, but during the school year.

So what I would propose here is that we have to go beyond using simply constructed response questions embedded in the on-demand assessment and move toward the development of performance tasks that would be administered during the routine school year.

Now my recommendation in this area would be to be very strategic. Use performance tasks to address standards not easily measured in the on-demand environment only. So we are not going to test anything with these performance tasks. We are going to address standards that just don't lend themselves to an on-demand environment. So we can think of scientific investigation, giving oral presentations, writing short research papers, so on and so forth.

Finally, teacher involvement in item development and scoring, to the extent feasible, is also critical to ensuring the quality, transparency, and integrity of the

system. It has certainly been our experience here in Massachusetts, where we try to involve teachers in many different phases of the program, that they end up being the best ambassadors for your program as it is being criticized by various people who like to criticize this sort of thing. So teacher involvement is also a critical consideration.

The last, in this final slide, I would just like to identify one of the questions that we were asked to answer was: what kind of local activities would be essential to implement the future assessment program? I would just like to mention a few ideas here at this time.

Firstly, we need tools and training that would help districts with curriculum development. Curriculum development is critical, but in most states it has been left largely to individual districts to carry out. Unfortunately, many districts lack capacities and expertise in this area, and could use external assistance.

I don't know. Recently, I have learned that there are a number of software applications that are very suitable for this purpose, but few LEAs can afford to license them or buy them. This may be a role that the state could assist districts in accessing.

Secondly, training teachers in assessment literacy is also critical, if teachers are going to play an active role in assessment administration and scoring, if they are going to make effective use of results.

Thirdly, I would say working with local districts to develop virtual, as well as face-to-face, professional networks for sharing best practices, including exemplar curricula, exemplar lesson plans, course syllabi, formative assessments, and benchmark student work.

And finally, states should be working with LEAs to develop resources that would assist them in communicating effectively with parents and community leaders about the standards, curriculum, and assessment system.

We cannot underestimate the need to continually build and maintain public support for our reform efforts.

So I would just like to summarize my main points quickly here, and I say I went through this pretty quickly, but that is okay.

So, just in conclusion, we need a unified system of assessments in which different types of assessments are targeted to particular uses. Summative assessments should be used primarily for accountability purposes; benchmark and formative assessments for improving teaching and learning. Summative assessments can serve as the anchor for the unified system if we leverage the annual release of items from those assessments to build the benchmark and formative assessments.

Fourth, a focus on higher-order skills is key if our students are going to be internationally competitive.

And finally, teacher involvement in building LEA capacity is key to implementing the system and building ongoing acceptance and

support.

So, with that, I will take your questions.

MS. WEISS: Great.

Yes, whoever's got questions.

MS. WURTZEL: Thank you, Jeff, very much. It is a really helpful presentation.

So one of my questions is, you talk about summative assessments being an anchor of a system and released items being used as benchmark assessments throughout the year. But, at the same time, you talked about performance tasks being administered potentially at critical times throughout the year for particularly hard-to-measure standards.

So I would like to hear your thinking about the relationship between the performance tasks and benchmark assessments, and how you see the performance tasks rolling up or not into a summative assessment score.

MR. NELLHAUS: So the vision here is that the performance tasks would

potentially contribute to the overall score on the summative assessment. But I understand there are technical issues there around calibrating performance tasks from year to year, so we can maintain a standard. So we may need to come up with at least some creative ways of reporting themselves the performance tasks.

It could also be that the performance tasks are used as part of the benchmark assessment system as well, but at least my initial thinking is that they would be integrated into the summative assessment results.

MS. WEISS: Jacqueline?

MS. JONES: Yes, thank you.

I'm really interested in getting more discussion around the formative assessments and your definition, and maybe some examples of how you see formative assessments being used.

But if you could talk a little bit more about the creation of this item pool from

the summative assessments, that would be helpful.

MR. NELLHAUS: Yes. Well, I know I didn't spend a lot of time differentiating formative assessment from benchmark assessment, but I will just try to do that as quickly as possible.

But the benchmark assessment would be an assessment that would be probably administered anywhere from two to four times a year. They would actually be constructed by the state or by the consortium. There would be booklets. It would be a mini-summative assessment, if you will.

I don't see them being as long as the summative assessment. The summative assessment, because of accountability purposes, needing to be highly reliable, needing to take as many as two or three sessions to administer, you would think the benchmark assessment could be administered during a class period, but it would be an actual test form that schools would have.

Those forms, because of the continual release of items from the summative assessment, those forms could be continually constructed.

Again, because we are not attaching any accountability to the benchmark assessment, there's no reason for teachers to try to teach to the benchmark assessment. I think that is an important point to make. There is a concern of teachers teaching to the test, narrowing the curriculum. And we are seeing that with the high stakes attached to summative assessments.

You don't want to exacerbate that with the benchmark assessments. So you want to make sure those are low stakes, that teachers are going to be free to use those in a professional environment to improve curriculum instruction, to frame differentiating instruction for individual students, and so on and so forth.

The formative assessment is anything from a teacher in a classroom just stopping for a second and asking the class a

few questions to check for understanding. It could be as simple as that. It could also be drawing from the pool that is generated from the summative assessment to construct quizzes, tests, whatever.

So a formative assessment, I would say, is teacher-constructed. A summative assessment would be more formally constructed by the consortium or by the state.

MS. WHALEN: In your vision of the summative assessment system, would it look different at all by subject or by grades? Would you see this three, eight, and high school looking similar?

MR. NELLHAUS: Well, I think at the high schools -- I would say for three through eight, every test looks a little different because of the subject matter. But I think there could be some consistency.

I think the general design that I have outlined could be applied to all the tests, reading, writing, you know, mathematics, science. I think you could

generally use the same construction, the same design.

So what I am putting forward here is an assessment that has some multiple choice, some constructed response. At least the summative assessment would have multiple choice, constructed response, and performance tasks that would be curriculum-embedded, and the results of those performance tasks would be rolled up into the summative results.

MS. WURTZEL: Can I ask a follow-up question about the performance tasks? Then Scott had a question.

So you mentioned the importance of teacher involvement and teacher scoring. Could you just elaborate a little bit about why you think that is a critical element with performance tasks, and given the experiences you have had in Massachusetts, what do you see as the lessons you have learned about how to do that, and what would be the challenges?

MR. NELLHAUS: Well, I can't say we have a lot of experience with performance

tasks in Massachusetts. I guess the best example I could give, and it is something we are currently doing, is for our English proficiency assessment, the speaking and listening component is a curriculum-embedded assessment. Teachers are trained through videos to actually observe students speaking and listening. They are trained in the videos to use a rubric, to actually score how they are doing in that area. Those scores are included -- the teachers actually indicate the students' scores on the students' answer sheet to the reading and writing portion of the test. They get included in the summative assessment.

But I think in terms of scoring tests, this is where teachers begin to better understand the performance expectations. So the investment it requires, it is going to require an investment upfront, one, in training people to use the rubrics to score, and then it is going to require an additional cost to do some sort of quality control

auditing by school to see how the scoring is being done.

So there will be a cost to this, but the benefit is that teachers are going to be involved in the system, and they are going to benefit from the training and understanding what the expectations are. Because for many of our assessments, the real expectations -- I am talking about the performance expectations -- what it is you really need to demonstrate to be proficient, is not very concrete. This is one way to make it concrete.

MS. WEISS: Scott?

MR. MARION: Yes, Jeff, just a quick I think clarifying question. I think it is an important point to bring out.

So you are distinguishing between these embedded performance tasks that are through the year that will eventually roll up as part of the summative and these interim little testlets --

MR. NELLHAUS: Benchmarks.

MR. MARION: -- or benchmarks that

are protected from accountability or separated from accountability; whereas the performance task -- so I don't see that much of a distinction there because I am hoping if these performance tasks happen during the year, teachers will use those results to either modify instruction or programs, or things like that.

I guess I'm pushing, why keep, if it is a good benchmark or a good interim assessment, why keep it out of accountability, so people attend to it? I understand not overemphasizing accountability.

MR. NELLHAUS: Yes. There is nothing to say that the summative assessments can't inform curriculum instruction, teaching and learning, as well as the benchmark. So I think, even though the performance tasks may be part of the summative, I'm associating with that certainly they can be used to improve teaching and learning. Okay. There's nothing to say that.

I don't have any research to base

this on, but my contention is, if we attach stakes to benchmark assessment, we are going to undermine the utility of those assessments. I don't know if there is any research that --

MS. WEISS: There is, actually.

MR. NELLHAUS: Okay. All right. There is some research.

MS. WEISS: CALC has done a bunch of research on this in healthcare to support what you are saying.

MR. NELLHAUS: All right. So I'll find that research sometime.

(Laughter.)

It's my gut right now.

(Laughter.)

MS. WEISS: I will send you the papers to support your gut.

MR. NELLHAUS: Okay.

MS. WEISS: Yes, go ahead, Henry.

MR. BRAUN: I have a question. One of the concerns I have is that, as we improve the quality of our assessments and make them more comprehensive and appropriate, it places

a greater burden, I think, on the teacher, as the instructor. So one of the unintended consequences could be a widening of the achievement gap because we know that, typically, disadvantaged students are exposed to more poorer-quality teaching. So, as we improve our assessments, if we are not improving the quality of teaching, righting the balance --

MR. NELLHAUS: Yes.

MR. BRAUN: -- we could, in fact, be increasing the achievement gap.

MR. NELLHAUS: Right. So I think it is safe to say that this is going to require, you know, moving in this direction is going to require a lot of professional development and capacity-building at the local level. I think, with that, we can close the achievement gaps. But I think you are right, that we can't do this and underestimate the level of effort that it is going to require.

MR. BRAUN: Thank you.

MS. WEISS: Any other questions?

(No response.)

All right, let's move on.

MR. NELLHAUS: Thank you.

MS. WEISS: So we have got, in our notecard system that is working here, we did get a request for each of you, when you ask a question, to make sure you identify yourselves because people can't see your name tags.

Jim, why don't you start by doing a quick who you are before you launch into what you have to say to us?

MR. DUECK: Right. Thank you.

Good morning. My name is Jim Dueck. I'm from the Province of Alberta, which is just above the State of Montana, and would be known by its largest centers of Edmonton, home of the Oilers and Mark Messier and Wayne Gretzky, if you know those names --

(Laughter.)

-- and also the Olympic city of Calgary, which also, by the way, today has 75,000 U.S. citizens walking its streets on a regular basis.

So it is my opportunity to give you a little bit of information with respect to the Alberta Assessment Program, as soon as I get that up on my screen, and indicate that we have more than 100 years of experience in the world of assessment.

And there is a huge difference in culture between the United States and Canada, and specifically Alberta. The cultural difference is this: that in the United States you have laws which govern assessment, which require assessment; in my part of the world, it is all through moral suasion.

That is significant because, therefore, you have to do a lot of support-building at the grassroots level, teacher involvement. So one of the themes you will pick up in my presentation is the necessity for ongoing teacher involvement as a way of promoting professional development.

Indeed, recently, or not recently, I could say 1993, the Congress of the United States reviewed our testing program, and their

particular finding had to do with extensive involvement because they understood that involvement has a way of improving classroom practice, both in terms of knowing what the standards are that are being required within the curriculum and also in developing their own classroom assessments, so that they will be able to be more expert in their practice. Last year, we involved more than 7500 teachers within our Province in some aspect of the assessment program.

So that you have some idea as to what our assessment program looks like, we test in grades three, six, and nine. These are summative tests. These particular tests in grade three are only in language arts and math. However, in six and nine, it is the four core subjects. Then, of course, grade 12, we have the diploma examination program, which is a high-stakes assessment program, not as high as what you would necessarily term "high stakes", but these are certification examinations for moving on. Therefore, they

count 50 percent of the mark that the student earns.

We operate with two streams at the grade 12 diploma level. We don't have, of course, any streaming that takes place in grades one through nine.

When I take a look at some of the questions that have been posed of this panel with respect to assessment, I encourage you to move as quickly as possible into the four core areas. I believe that it maintains the emphasis in all subjects, and therefore, does not show any subject to be of less value.

I encourage testing annually at every grade because of the distributed accountability that comes about, the value-added measurement for individual students as well cohort comparisons that are then possible.

I encourage with respect to an exit examination at the grade 12 level because it motivates students right to the very end of their school and career, the K-to-12 school

and career, and it also provides fairness to students in terms of the consistent interpretation when selections are being made for post-high school involvement.

With respect to the kind of a vehicle that is used, I am very intrigued by computer-adaptive assessment and see it as a preferred future in the world of assessment.

When we take a look at our grade 12 assessment program, as I have said, it is an examination that requires or that constitutes 50 percent of the student mark. The classroom experience is 50 percent, and we utilize a number of activities, the projects, the labs, and so on, which constitute that 50 percent of the mark.

And then the central marking system is a very key aspect of teacher involvement. That is where we are, then, able to mark the machine-scorable portion as well as the open-ended responses, which constitute the other 50 percent of the mark.

I need to put an elephant on the

table. The elephant on the table is that there is a myth regarding the ability of multiple choice questions to go at the higher-order thinking skills, or the HOTS, if you like.

We know that when we have seen many kinds of multiple choice questions, they tend to focus in on the LOTS, the lower-order thinking skills, because they are factoids; they are recall, isolated facts. There is very little of an opportunity, then, to go into what I would refer to as the HOTS.

Another example that has been used, and I know that I am sure that this conjures up, because I see people's heads moving up and down, I don't need to say anything else about the factoids, but I want to show you what I believe is a very well-constructed, machine-scorable multiple choice question.

No. 1 deals with commonality. There are two distinct, but related, events that are being assessed.

No. 2 deals with cause and effect,

whereby the student is looking at synthesizing knowledge, and we also now are incorporating numeric response, whereby the answer is not even located on the paper, as another way of assessing the students.

If I move into more of an example within a subject, it is our intent within our exams to use real-life context as much as possible. A reader of the examination who had not taken the course would actually learn about the real world. So our emphasis here is to try to embed science into technology and society. Therefore, this particular item will get at the whole issue of how a drug can contribute to autism, for example.

If I move into another area of what we try to do in enhancing our examinations, you will see that in the very bottom, in order to try to get quality as much as possible, we will actually quote sources, providing copyright, and so on, so that we are able to use research, charts, pictures, and every way possible to hook the student into the

examination question and provide a much higher level of quality.

As I said, we are looking at trying to get both the science as well as the knowledge in society. The questions that are in the red ink, and this particular chart looks funny because it doesn't come out the same as on mine. I'm not sure what color of lightbulb we have in here.

But, at any rate, the top part of the questions, which are the true red questions, they are dealing with the science aspect. They are looking at the pure scientific knowledge in the particular course. If you go to the bottom portion, what looks brown on yours is that it is knowledge in society. We are trying to pull out of the student their understanding of the relationship that issues will have as well as with matters of government policy, as well as with matters of finance and cost implications, and so on.

Moving further into the biology

example, this is a rubric. This particular rubric just deals with the area of science. So this is the pure knowledge assessment of the biology question.

And if we move into the next one, we move into the technology and society aspect, and the questions that would come from that particular area.

So these are the "look-for's" that teachers will use as they come to the marking center in order to mark the exams as they arrive.

Moving into pure math or mathematics, we have, as I said, two streams: the pure math, and there is only one way of solving in the pure math. Hence, the question will deal with algebraically. Then we have the applied math, whereby students are able to move into multiple strategies in order to come up with the answer.

Now what I wanted to indicate here is that these exam questions are all on our website. So that students are able to

frequent this. Teachers are able to frequent this on an ongoing basis. They are able to see the rubrics.

As a result, of course, we have to release items annually. We try to release 20 percent of our diploma items annually, and our provincial achievement tests we release in their entirety every two years.

I forgot to indicate that our Province makes 102 tests every single year. So, therefore, we want to try to overcome the issues of security and the like, and, of course, it takes a great deal of work.

One clarification before I move on. We are faced with budget problems, just like everybody else. We have been doing studies of our machine-scorable portion, our multiple choice questions.

We find that there is a correlation of .95 to .98 on the machine-scorable portions with the exam result in its entirety. As a result of that, government has made the decision to remove for the time-being the

math/science written component because the correlations are so high.

Moving into another area, source-based questions, which are things that we were inventing in the Province of Alberta about 25 years ago, and source-based questions allow us, then, to do an analysis of one picture through a machine-scorable response.

Then, moving into the second particular picture, and then the third one, if I can get that up, the third one, which really is the methodology that I wanted to try to point out. That is, that as a principle, the knowledge of topics in a syllabus taken at different times of the year's study to answer a question never asked before is what we are really trying to do. That is where we are able to get at the high-order thinking that we would like to see within our student population.

We provide 11 different types of accommodations for students who have special needs or who are ESL, as we have many of those

now coming into the Province as well. We essentially provide all of the accommodations that the schools are able to access during the year, so that the student doesn't have any different set of accommodations to be thinking about when they come to the examination.

Now let me move into the area of teacher collaboration specifically. With, as I said, moral suasion being the whole issue that we are dealing with on an ongoing basis, we want to make sure that we are able to provide the very best professional development that happens, and our teachers are telling us constantly that their involvement in some aspect of the assessment program does just that for them. So we encourage teacher involvement in all phases.

There is the test preparation aspect, which gives people an idea of how they can go about constructing their own tests once they are involved in their own classroom assessment activities. During this time, they can sit with an expert from the Department who

does nothing but eat, drink, sleep, dream about how they can make the most proficient examination question in the world, and they do it over the entire year.

They are also involved in the reviewer phase, during which time they have to sign confidentiality agreements. There is a bit of a security risk, but we have only had one breach of that security over the many years in which we have been involved.

There's also the field-testing phase, when we test the questions as to their reliability and their validity, and teachers give us information back on all of the test questions before they get used. We administer these particular questions in all segments of the Province, so that we have an idea as to what the response would be like from the various groups within the population.

Papers are then selected for training that we will do with markers, once they are brought back into the marking center. Reliability reviews are undertaken. Then,

also, examples of different scores on the rubric are also identified. So, usually, we pull about 1,000 tests before the marking exercise actually begins that we can use in all these various phases of review.

The markers, as they then come to the marking center, and the group leaders will meet first at the top little table for a few days to go through all of the assessments. Then they will each chair a table of markers during the actual marking event which follows. In a sense, therefore, we are able to capture in on distributed leadership across the Province when it comes to the issue of assessment.

Our reliability reviews really have as the theme fairness to students. That is what it is all about. We want to make sure that the marking is as consistent as possible. So, therefore, the reliability reviews involve twice daily the teachers in an exercise whereby they take a look at papers. This is sort of the one reliability review

that goes on.

The red shows in the particular question that on this question 12 teachers gave it an excellent, 44 gave it a proficient, and so on.

After having, then, a discussion, following what their original assessment was, we, then, get into the reviewing of it. You can see the shift in where the teachers have come in their thinking, and the response, then, shows that the proficient is the one that has been most selected.

From there, we also provide teachers individually with the result of their reviews. So this particular teacher, number of papers marked, is marking many more than the mean.

With respect to the effectiveness, the marks are lower than what the mean would be. Then, with respect to the third reader rate, they are having many more third reads applied to the marking. So, therefore, there is a great deal of need for this particular

teacher to have some coaching by the exam manager in order to help that person more within their task.

And if we take a look to see whether it was just one item of the question that was causing problems, in this particular case, it shows that every single question was, indeed, problematic for this teacher. They were always on the low side. So, therefore, it is very pointed as to what kinds of coaching that a person can provide for that particular teacher.

We go into the third reader rate once the exams have all been marked twice. Then the teachers who we want to hold back as being the most proficient go through the exercise of the third reader rate. Once our examinations are completed, we work hard, then, to work with our advisory committee, stakeholders, and professional people as to what we can do and learn in order to make our examination questions better the next year.

We work and make ourselves

available as much as we can within our budget to provide professional development at the schools across the Province, and our staffers are able to provide about 50 of those during the year. Then we also sit down and work with our pre-service institutions in finding out how we can assist them in preparing the teachers of the future.

Our theme in all of this really is as follows:

That we don't want to be a mirror. We want to be a prism -- I didn't say, "prison"; I said, "prism" -- when we are constructing our test questions. A mirror merely reflects the input and is lower-order thinking in nature. A prism takes the knowledge and breaks it apart, so that synthesis can, then, be undertaken and, therefore, demonstrate the HOTS much more than the LOTS. That is really what we are trying to do with our examination questions.

Now all of this does tie into our provincial accountability system. First of

all, you will notice that on our accountability system our assessment program is not the entire system. The assessment program I have been describing are the areas right here, the provincial achievement tests and the diploma examinations. That is our assessment program using paper-and-pencil tests, or we are now moving into technology.

We have many other measures that we incorporate on building our report card, and this is a provincial report card. Every school jurisdiction and every school receives a similar report card on their performance, using a wide variety of measures. So it brings in all aspects of the school's delivery, including the quality, the involvement of people in the school, making sure that we have completion rates and the dropout rate issues; also, the matter of safe and caring, and all those issues that are important, an assessment of art and music and phys ed. All of that constitutes a way of taking a look at the system.

We don't only measure it on the basis of raw score. That is a fallacious issue because raw score is definitely impacted by socioeconomic status. So, therefore, we have to neutralize for that.

So, while we will evaluate on the basis of the raw score, which is the achievement column relative to standards, we also do an evaluation with respect to the improvement, and whether or not there has been significant improvement that has occurred, because that really is the leadership index, leaving a situation better than when you found it. So that is the kind of report card that we have constructed, of which the assessment is one portion of it.

So thank you very much for this opportunity. I trust you have some opportunity to understand. I have gone too quickly because I have made it by 10 seconds.

(Laughter.)

Thank you very much.

MS. WEISS: No, I think you have

definitely set a new bar for the number of words in 20 minutes.

(Laughter.)

I hope the transcriber's fingers aren't falling off over there.

So, Jim, let me just kick it off by asking one question. Over the course of time, as you have been doing these kinds of assessments, have you seen, are there correlations with what's happened with student achievement in the Province, and how do those match the different evolutions that you have seen as you have changed your procedures over the last --

MR. DUECK: Yes. We participate in all of the international assessments, the PISA, the TIMMSS, as well as our own assessment called Pan Canadian Assessment Program.

If I were to go back in time about 20 years, the Province would be referenced probably by the word "middling". Over the course of time, once we introduced our

provincial achievement tests in the early 1990s, the Province was able to identify significant weaknesses in its curriculum, ratchet-up the standards in that curriculum. In the 2003 PISA results, Alberta was the highest-performing country because we oversample and qualify for a country, the highest-performing country in the world. In 2006, we dropped to second, behind Finland.

MS. WEISS: Do you think, I mean, can you tie it back to any of the changes and things you have been doing on the assessment front? Has the assessment system changed over this time?

MR. DUECK: I am just going to give one other response, I think, to the earlier part of your question.

MS. WEISS: Okay.

MR. DUECK: That is that what we have found in the use of our testing program is that people, therefore, are becoming very familiar with what the standards are for their particular area of the curriculum.

What that does is, then, controls the tendencies toward great inflation, which is often a very subtle and significant way of reducing student achievement. We have found that, as grades are being inflated, that student achievement on our assessments goes down.

So, therefore, we want to make sure that our teachers are aware of what the standards are, and our assessment program has provided that professional development to help them know those standards and apply them to their students.

Does that answer your question?

MS. WEISS: Yes. Thank you.

Questions?

MS. JONES: You have described teacher participation in the design of these assessments. Can you speak to whether or not there's a relationship between that participation and your own design and implementation of assessments in their classrooms?

MR. DUECK: Yes. In our Pan Canadian Assessment Program, which is like the other assessments that has a survey that goes along with it as well as the actual test, over the course of the years, and every year, Alberta teachers have demonstrated that they are the least likely in Canada to utilize non-academic variables when they assess marks to students.

Whereas, there is a tendency from teachers to incorporate behavior, attendance, other kinds of student compliant issues, they are very committed, because they are showing that every year the Alberta teachers are the least likely to use those variables. To focus in only on the students' demonstration of the learning, that is the key for us.

So we feel that, by giving the teachers that constant feedback in the three, six, nine, that teachers, therefore, are very much focused on what the standards are, rather than judging students on compliant behavior.

MS. WURTZEL: So I have a very

practical follow-up question, which is, could you speak to the percent of teachers that might be involved in these activities in any given year, and how much teacher time is devoted to that?

MR. DUECK: Sure. We have in our Province 35,000 teachers. As I said, we involved about 7500 during the course of the year. Now there is overlap. Some teachers are involved in more than one activity, and I can't go into any greater specificity as to the number that have been involved. So that is one aspect of it.

Then we also are heavily engaged, as I am sure you are over here, on the whole issue of assessment for learning. We have now taken that role within the Department as well to work in the assessment for learning because we have been involved already in assessment-for-learning kinds of testing over the course of the years. We are now moving into a new generation, utilizing technology, and so are heavily engaged right now in developing a math

diagnostic assessment.

MS. WEISS: Other questions? Gary? And start by introducing yourself.

MR. COOK: Gary Cook, Wisconsin Center for Education Research.

I guess I have two questions. One is, what is it about Alberta teachers that has them focus on the content more than other Canadian teachers? And maybe I misunderstood you, but because of the relationship between the constructed response items and your multiple choice items, on the summative assessments to remove the constructed response, what kind of consequences do you think have occurred as a result of that? And is that, if they are not negative, is that a result of your teacher education and the relationship with IAGs?

MR. DUECK: Okay. First of all, with respect to Alberta teachers and their focus on content, you probably are familiar with the research that shows that, when you institute an assessment program, results will

immediately spike upward. We experienced that. We would probably be like most jurisdictions and find out that what that was doing was causing people to teach the provincial curriculum, and not just the one that they wanted to teach. So we saw that take place.

Then, of course, you also may have seen a plateauing of results that follows. So what we have been working on now is another level of accountability, which is the report card that you see up on the screen, as reminding us that you have to be focused in on improvement. Every situation can be improved through professional development, resources, whatever the case may be.

So you can see, if you take a look at the provincial achievement tests this past year, that has improved significantly. We are very pleased to see that that kind of result has taken place.

With respect to the constructed response, that is something that the

government has just determined in the last few weeks. There is always the concern that people have that, can you really mark partial responses when you use that approach? We view this as a summative test. We can get at partial approaches using numeric response. We are able to assess partial correct response using that idea. But we will see how it goes as the year progresses.

We have now gone through the phase of making announcements. Of course, people are always suspicious, but, as people understand the high correlation that exists, that has muted some of the opposition.

MS. WEISS: Other questions? Yes, go ahead, Henry.

MR. BRAUN: Henry Braun, Boston College.

Jim, can you say something about, over the last, say, 10 or 15 years, changes in the requirements for teacher credentialing in Alberta? Have they gone up or stayed the same? What have you seen in terms of

attrition rates, teachers leaving the profession?

MR. DUECK: The credentialing requirements have remained the same over the course of time. I suppose that is about all I can really say, is that they have not changed at all over the last several decades.

With respect to attrition, I would make these observations: we have had a significant attrition of teachers in the last five years. We have sort of that baby-boom element come through. So we have had a definite loss.

What happened as a result of that is that we were bringing in teachers from many other parts of the country in order to replace the workforce. As I said to you a little earlier, our research across Canada showed that there were tremendous variations on teachers using non-academic factors in assigning marks.

So, therefore, in the most recent survey of that, we found that we had moved up

the ladder in the wrong direction quite significantly. So that has been a concern to us, and we need to work hard at professional development of our new cast of teachers.

MR. BRAUN: Thank you.

MS. WHALEN: Can I ask a quick question?

MS. WEISS: Yes, go ahead.

MS. WHALEN: You only test in three, six, nine, and then 12th grade. Have you seen any unintended consequences of pulling out certain benchmark rates that are associated with these higher-stakes-scored skills --

MR. DUECK: Yes.

MS. WHALEN: -- and the quality of teaching and learning, then, in those classrooms as well?

MR. DUECK: That is the reason for one of my recommendations for your consideration. Because, very definitely, if you only utilize benchmark assessment, what ends up happening is that the grade four/five

teachers do not experience the same ownership as does the grade six teacher.

I recall, one of my first days as a superintendent, the grade 12 teachers coming to me and suggesting, because at that time we only did grade 12 testing, that we really ought to bring it down to grade 10 and 11 because, after all, they needed to share the ownership of all that was involved in that.

In our own assessment of results, the highest grades in senior high are grade 10. The grades drop in grade 11. They drop significantly in grade 12, and the diploma exam brings a further reduction significantly.

MS. WEISS: Yes, quick, the last question.

MR. NELLHAUS: So what I understand is you release your tests every year, correct?

MR. DUECK: A portion of them.

MR. NELLHAUS: A portion? Oh, a portion?

MR. DUECK: Yes. The grade diplomas, 20 percent; PATs, every two years.

MR. NELLHAUS: Okay. So, just if you can answer this quickly, how do you equate the tests from year to year?

MR. DUECK: I am a policy person, so I would have to bring my technicians in to do that.

MR. NELLHAUS: But you do equate?

MR. DUECK: Oh, definitely.

MR. NELLHAUS: Okay.

MR. DUECK: We definitely go through all of the world-renown equating processes.

(Laughter.)

MS. WEISS: Thank you. That is what we will be using, known as world-renown equating processes.

(Laughter.)

Let's turn to Laurie Wise as our next speaker.

We need to pass the clicker on down.

And, Laurie, why don't you start by introducing yourself, and then we can dive in?

MR. WISE: Hi. I am Laurie Wise. I am principal scientist with HumRRO.

I have been active in a number of efforts that the National Academy of Sciences and other groups have had to look at testing and test use policies. So the Board on Testing Assessment, and I am currently involved in working with eight-year APA and NCME on the revision of the standards for testing.

I come to this somewhat more as a psychometrician in comparison to the first two speakers, and not as a person who actually owns or has had to run an assessment. But I will say that the one thing I have learned dramatically, and, unfortunately, over and over again, in this business is that content matters. So I am interested in thinking through the implications for content of this new opportunity both of the common core standards, and then of this program to help build assessments around them.

Because I can't remember what state

I am in or what day I am in, I put an advanced organizer out here of some of the topics that I would like to cover as we think about how to build this program of common assessment.

So, first and foremost, and the thing if you read the validity books, you start with the purpose. What is it you are trying to accomplish? There are many listed, among them including school accountability, identifying high- and low-performing schools, teacher or principal performance indicators. I won't get into the politics of evaluating teachers on this, but it does provide useful feedback and guidance that can help teachers and principals improve and can hold people at some level accountable for the effectiveness of instruction.

Evaluating instructional programs and really improving the programs is an important component of what we are trying to do, I believe, not just measuring their effectiveness. And finally, providing diagnostic information about individual

student deficiencies.

The second topic is test content. There are questions about the extent to which we focus on current grade-level expectations versus talking about vertical alignment of expectations and building tests that sort of measure accumulatively where students are on that whole pathway from entry into school to readiness to get out of school.

Test administration. We have already had some discussion about, is it one big summative test at the end of the year? Are there more opportunities for students to take the assessments during the school year?

Test format is another issue. It can range from large, group-administered, paper-and-pencil tests, computer-based, computer-based-adaptive tests, and then assessments with significant open-ended questions that can't be machine-scored, which sort of works against the adaptive or sometimes even computer-based advantages.

And finally, I want to end with

talking a little bit about sort of what kind of validity evidence do we need to support these assessment.

So let me say at the outset I think we have a tremendous opportunity here to really move things forward. As I understand it, my goal or my task is not to look at the current assessments systems and say, "Oh, why?", but to look at the assessment systems as they might be and say, "Well, why not?", if I can paraphrase. You know I didn't make that up.

(Laughter.)

So, first, I think we have to begin by acknowledging that one test is unlikely to meet all the purposes well. Developing tests that are deep enough to be diagnostic and broad enough to be summative over a large domain, especially a cumulative domain of accomplishment, is a well-nigh impossible task to have it all at once.

An example that I think is very concrete is with NAEP. NAEP uses matrix

sampling. It is great for getting aggregate-level information, how are we doing as a state and as a country, but it gives very little information about how individual schools are doing, and nothing really about how individual students are doing.

So a goal, as we look at this redesign, I think will be to see if there are ways that we can build assessment systems that may be more than a single test, that can support the improvement of instruction as well as holding students and teachers and schools accountable for the results of instruction.

So we are about to see, and I have to admit, there have been several people inside the tent that have seen the common core standards grade three or K through 12. I have only seen, myself, so far, the readiness standards, and I am anxious to see what they look like.

The hope is and the advertisement is that the new common core standards will be fewer, more focused, will be clearer, and will

be higher in comparison to what most states have now on their plates. We'll see.

Also, there is an attempt, then, to line up grade to grade in a pathway that leads to the readiness standards at the end. That would be a significant improvement over what we see in many states, where the content standards for each grade were developed somewhat independently of each other and sometimes with some levels of articulation, but not always with a real clarity of what more we want students at this grade to know that they didn't know at the prior grades.

What I think would be useful is, with a common set of standards, we have an opportunity to really study how to teach them in-depth that we don't now. I am reminded, when I was really young, we didn't have any standards to speak of, and test publishers built scope and sequence charts. They decided sort of what is the best way to teach things. It is hard for publishers to do that when they have 50 different sets of content

standards to meet.

So the hope is that we can go back to thinking about intelligent sequencing of the material within year as well as the cross-year content standards.

Now I want to talk a little bit about the reporting scale. We have seen quite an evolution in reporting scales. There has been a movement to reporting performance levels, to making some judgment about what is satisfactory performance and then reporting what percentage of students are at that level.

That doesn't by itself tell you very much about what the students actually know and can do. As we have seen in many states, the performance standards are somewhat uneven across grades. So a student may be up one year and down another. There are efforts underway to articulate not just the content standards, but the performance standards across grades.

So the alternative is that we often report things on a reporting scale. The

states and their vendors have great fun inventing these reporting scales. So it has gone up from 201 average score to an average score of 205. That is great, but what does that mean? Because the scale by itself doesn't mean anything.

So what I would be hopeful for and would be encouraging of the consortia as they move together to build these tests is to spend some time and effort identifying, developing, and thinking about a learning trajectory scale; that is, a sequencing of content within as well as across years, and being able to report where students are in mastering that sequence of content, rather than just simply either of the two somewhat arbitrary reporting scales that we have been using now.

So I am also hoping that the test systems that are developed will have both the formative and a summative component. The formative component would include multiple assessments focused on smaller sets of content.

One of the problems now is that we try to make the summative assessments diagnostic, but they cover such a broad range that you are ending up with only a few questions for each content area, and you don't really have reliable and independently-valid information about mastery of specific content areas.

So, if we had multiple assessments during the year and if they focused on the content that was sequenced for that portion of the year, we could have much greater depth, and then some way of summing them up across the whole year to get coverage of the whole domain.

The model that I had in mind is to take advantage of computer administration, computer scoring, to use adaptive testing that focuses in very quickly on where the student is in this learning sequence, and then ask a significant number of questions that really get at the level that the student is now and needs to move in next, and not too many

questions that are way below or way above that student's level of mastery.

And the advantage of adaptive tests is that you are scoring as you go along. The student hits Enter and is done, and you have the score right there. You can provide immediate feedback to the teachers, to the students themselves, and so on.

But I would see this coupled with, at the end of this short, objectively, or machine-scored section, opening up some deeper open-ended questions that really provide diagnostic information that let you see more clearly why the student doesn't understand something or what it is about the skill that they haven't really mastered.

Then allow these open-ended responses to be separately scored afterwards. Involve the teachers in the scoring, maybe with some external audits. Provide much deeper diagnostic information about what the students know and can do. Use the additional information to confirm or, in some cases,

modify the initial score-level estimates.

We are constantly railing against not a single source of information, and here is a way of figuring out how to use the deeper diagnostic information in a way that could also correct situations where the initial summary score may not be the best judgment about where the students are in mastery of the content.

Then for the summative component, I think it is important that we could look at within-year growth by testing multiple times. So we could see progress from the initial to the final assessments, which is a much clearer assessment or indicator of the contribution of the teacher and the curriculum at that grade level than just looking at the final outcome, which may vary a lot as a function of where the student was to begin with.

Seeing the actual student progression, both for the students and parents, as well as for the teachers, I think gives a better and clearer and more dramatic

picture of what the contribution of instruction has been during a particular school year. It is certainly a better way to assess program and teacher effectiveness, if that is what we want, to the extent that that is an important focus of these tests.

We do need an end-of-the-year estimate of the overall progress toward readiness by the end of high school. We need to map the within-year trajectories onto larger models of the K-12 learning process.

We do need to report the absolute level of where students are at the end of each year to their parents and to the students themselves, and ask, is the student on track to being ready by the end of 12th grade?

In addition, we would also still want to look at year-to-year growth from the summative end-of-year measures as an alternative or additional way of providing information for program and teacher accountability.

Right now, and I work with several

states, I know how strapped they are in terms of resources, and especially time, just to get the programs up and running. While most of the states really, frankly, have complained a bit about the peer-review requirements -- (laughter) -- there have been a number of very positive, I think, outcomes. It has provided an opportunity for discussion of the states and their vendors and interested parties, to really think about what kinds of validity evidence are useful to collect, and how this validity evidence could be used in an ongoing process improvement to continue to improve both the assessments and the use of the assessments.

So what kinds of validity evidence might there be with this somewhat more integrated system that I just described very briefly?

Well, first of all, we would want to look at some issues of internal consistency across the items assessing specific content clusters. Right now, in the summative

assessments, there aren't enough of them to do that well. Many of the benchmark assessments are really full-range summative assessments, just given at different times of the year.

By having these interim assessment that are focused on a portion, a smaller portion, of the content, of the objectives for a given year, we should be able to do more with psychometrics to assure that the questions targeted to specific clusters of content objectives actually have some coherence.

We also need to establish evidence for the mapping that assures coverage of all of the objectives at appropriate depths of knowledge. So the kinds of validation we do now for our assessments, which are mostly used to indicate whether the student has mastered the material for his or her grade, is an alignment study that says, are you covering the content with the test questions, and are you covering at the right level of depth? And we need to continue to do that.

In addition, it would be useful to also collect data on, are there common patterns of progression? Are we teaching the material in a logical sequence? Are there alternative sequencings that might be more effective?

In addition, now we are wanting to look at convergent validity studies. I think it is always useful to go back and gather teacher judgments about where the students really are and say, do the tests provide similar information and, if not, why not? What is it about either the teacher judgments or the tests that is giving you a different picture of where some or many students are?

We can now, I think, do this with respect to the learning trajectories and ask the teacher where along the sequence of mastery of material each individual student is and compare that with the assessment results.

Then, finally, we have ignored sort of predictive validity studies for a long time, but now if we are back in the game of

building up to readiness by the end of 12th grade, I think it is on the table to look at evidence that the student mastery of material at one grade predicts their ability and their likelihood of going ahead to master the material at the following grades and reach readiness, desired readiness levels by the end of 12th grade.

So just a few recommendations. Again, the Department, I think, has a major challenge in trying to figure out how to manage this procurement. And I wouldn't begin to be expert on many of the things, like how the states build the consortium, et cetera.

But one thing is it does seem useful to not put all your eggs in a single basket, to fund several different consortia with multiple or different approaches, that would allow you to take some risks. If you just fund one thing, you can't afford to do anything too innovative because it might not work, and then you would have little.

If you fund several things, you

should be able to take some risks. I have in mind great improvements in assessment systems, but you also need to be constrained in that you have to be sure that at least some of them will, in fact, work out. So take some risks, but not too many.

Support analysis of the sequencing of contents and the building of learning trajectories. It is not just about building test questions. It is about the whole model of how the content fits together, is mastered, is taught, and is assessed.

Emphasize possibilities that do provide diagnostic information as well as summative information. Look for these integrated systems. Multiple focused assessments each year. Quick turnaround scores coupled with hand-scored responses to provide more diagnostic information, as opposed to simply summative end-of-year tests, followed by benchmarks that states buy on their own that may not be well-connected.

Emphasize teacher involvement. I

think that is a good thing. You have seen examples from Alberta about how that is accomplished.

Do release items, but, of course, equating requires that you keep some items secure and are able to re-administer, so that you can gauge the comparability of test forms from one year to the next.

It might be possible to release all the open-ended items, which I think provide the really rich diagnostic information and encourage deeper teaching.

And I am done.

MS. WEISS: Thank you. Thanks. That was terrific.

Let me start by asking you a question about the notion that you might use computer-adaptive testing to sort of hone in on the areas that you want information on and then give open-ended questions about those.

So that means, presumably, that in a classroom -- and this isn't necessarily a bad thing, so I am asking a logistical

question as much as anything -- that in a classroom, a teacher might have a whole bunch of different open-ended assessments that they are giving to different students, based on the results of the computer-adaptive test, if I am understanding what you are suggesting properly. I am just wondering about the --

MR. WISE: One question -- and I am anxious to hear the technology panel tomorrow -- but one question is, can the open-ended questions also be computer-administered, so the computer can quickly find --

MS. WEISS: The right thing and pick --

MR. WISE: And at the higher grades, I am pretty sure that is quite feasible. In fact, I think if you ask most students at the higher grades to handwrite answers, they are at a disadvantage because they don't do that as much anymore. I would be at a disadvantage if I had to write a long essay.

So just because it is broad or

open-ended doesn't mean that it is paper and pencil. In fact, I think we will hear probably tomorrow in the technology panel about some innovative, more scenario-based ways of collecting evidence about inquiry skills and other important skill areas that aren't just essay questions.

MS. WEISS: Uh-hum. Great. Thanks.

Let me turn to my colleagues. Do you guys have questions?

MS. WHALEN: So I just had a clarifying question. In one of your slides, you talked about how formative assessments would be administered throughout the year, and then in another slide you talked about being able to measure a student's progress throughout the year, from the beginning of the year to the end of the year.

Do you envision having that be part of the summative roll-up score, where I am assuming there would have to be a formative assessment in the beginning of the year? Or

were you thinking of just two types of assessments that would be used to roll up to the summative score at the end with formative assessments in between? Or did I misunderstand?

MR. WISE: Well, no, I was purposely vague because --

(Laughter.)

MS. WHALEN: I wrote down the same question. So I am glad to hear that.

MR. WISE: I wanted to leave a lot of room for creativity among the states and their vendors.

(Laughter.)

However, what I would suggest as one possible model would be maybe there's like quarterly assessments at three points during the year on the summative, and the quarterly assessments don't cover everything, but they cover the material that has been taught to that point.

Then you look at mastery of that material. Then maybe at the end there is a

way of really using this information. So, if they have mastered the material in the first quarterly assessment, maybe you don't need to retest that part again. Or there may be more creative ways of integrating the results from the quarterly or interim assessments with the results from the final, end-of-year assessment.

MS. WURTZEL: Laurie, I would like to ask you a question about the issue of learning trajectories that you raised and reporting on a trajectory scale. So can you talk a little bit about what you see as sort of the opportunities and challenges in creating learning trajectories grade by grade and then, also, within-grade learning trajectories, given that this might be an assessment used by multiple students with multiple curriculum?

MS. JONES: Are you suggesting grade by grade, talking about trajectories?

MR. WISE: Oh, yes. I really think that within-grade as well as across-grade, and

there is some research. There has been, you know, starting back with the National Academy, the National Research Council report, "Knowing What Students Know", and so on, there was discussion about learning trajectory, and there have been some efforts to try to define them.

I would make one important distinction. You can define learning trajectories that somehow there's evidence that you absolutely have to know this before you can know that. That is the reason for sequencing.

But you would also get learning trajectories based just on the order in which you teach things. I think the trajectories across grades really reflect judgments about where it is appropriate to learn different things. So that it makes sense, then, to say how far you have progressed, how many of those things you now know. Within-grade it is the same thing, if we pay attention to what is the optimal sequence of teaching the material.

So, rather than trying to say, on day one we are going to teach you all the content standards, and then on day two we will sort of fill in those that you didn't quite get on day one, we will start you with some portion of the curriculum that may or may not be a true prerequisite for the later curriculum, but will certainly, in some logical and pre-planned way, build as the year goes on.

MS. WEISS: So can I just follow up on that for one second? So I think what you might be saying, if I am understanding it, is that, if you have a computer-adaptive thing mediating in the middle of this, you could even have different states and different districts even teaching things in a different order because the computer is sort of matching these quarterly things to whatever order your scope and sequence says to do it in. Then the research looking at all of this data could start over time figuring out for us, is there an optimal trajectory hidden in here someplace

and help improve those trajectories over time? Am I understanding that right?

MR. WISE: Well, first of all, I am a researcher, so I support doing things different ways and encourage research on that.

(Laughter.)

But, certainly, there would be that capability.

I do think, as most states get together and look at this, they would be more comfortable with recommending a regular modal sequencing and having the tests reflect that.

MS. WEISS: Right.

MR. WISE: So that there is some comparability across schools and districts of the interim assessments as well as the final assessment.

MS. WEISS: Okay.

MR. WISE: But take some risks. You might want to try out letting different districts sequence things differently, and just be able to go into the computer and say, "We've covered objectives one, seven, and

twelve. Now build me an interim assessment that tells what students have learned on those topics."

MS. WEISS: Jeff?

MR. NELLHAUS: Just following up on this trajectory scale idea, it sounds to me -- I am trying to envision what it would look like. Would it be a reporting by standard perhaps, whether the student has mastered the standard or not? I am trying to understand exactly how this would be manifest. Is it more like competency-based testing, and you are really seeing whether or not students have mastered particular competencies at different points in the year?

So if you could just talk a little bit more about what the scale would give you?

MR. WISE: One sort of contingent piece of information is sort of, when NGA and Achieve, and so on, talk about fewer, clearer, higher, how many fewer are there going to be? Right now, most states have so many objectives that having any very valid and

reliable measure of mastery at the individual objective level would require that you spend more time testing than teaching. I think we don't want to do that.

So I am open to either having fewer key standards and assessing mastery standard by standard of those or of having clusters of standards that are not quite as broad as the content strands that we have now, but maybe not quite as fine-grained as the individual objectives that most states have, and then asking, to what extent has the student mastered this sort of related group of things that are taught somewhat together as a unit?

So I am being, again, a little vague because the devil is in the details, and the content people are going to have to really dig into this. It is going to require some development effort beyond just building the test questions.

MS. WEISS: Henry?

MR. BRAUN: Henry Braun.

As you probably know, in England

they have developed these learning trajectories and a set of up to 10 levels or milestones from the early grades to the later grades, but they are not particularly tied to a grade. So a student can be assessed relative to their progress, and it might be in the fourth grade, it could be in the sixth grade, and so on.

Do you have some thoughts on whether we should experiment with that sort of model?

MR. WISE: Well, you know, I didn't want to step on the political landmine of out-of-grade testing.

MS. WEISS: Although it was in our notice as a question.

MR. WISE: Yes.

MS. WEISS: So we opened the landmine for you. So feel free to walk in.

MR. WISE: But it does seem like, with an adaptive test, you can quickly determine that the student may not even be anywhere in the current year. If you have a

trajectory or a model that spans grades, it would be possible to go down and find where the student is in this sort of cumulative assessment of, from K to 12, how far toward readiness are they?

Then maybe so as to take appropriate action, if the student really is not at a level where they are likely to benefit from the instruction that they are currently getting.

MS. WEISS: Gary, did you have a question?

MR. COOK: No, I don't.

MS. WEISS: Any last questions?

MS. WHALEN: I just wondered if you wanted to go into a little more detail about the validity evidence.

MR. WISE: No.

MS. WHALEN: No?

(Laughter.)

MR. WISE: Well, no, there should be some, but that would be a very long discussion, and the yellow is about to go red.

I do have an opportunity tomorrow to talk in a little bit more focused way about high school. We will revisit that tomorrow.

MS. WEISS: Great. Thank you so much.

Gary, over to you. Gary Cook.

MR. COOK: I am Gary Cook from the Wisconsin Center for Education Research.

I want to thank Jeff and the Bostonians for their hospitality in having us here and for the Department of Education inviting us.

I guess my task is to chat about some ideas and thoughts about the assessment program that is associated with Race to the Top funding. I am not going to talk about the assessment guideline, guiding principles, because, actually, it has already been talked about.

I am coming at it from more a perspective of alignment standards, how to think about what this might be when we talk about a consortium of states doing the same

thing, and what is involved in that.

Then, particularly, I am coming at it from an English language learner perspective. So I am interested and I do research in the area of English language learners. So that is kind of my take on things.

So I am actually taking a chicken's way out. I am going at the 10,000-foot level and then the 10-inch level, and the 11-inch to 9,999 I am going to leave for the people who are really technically-brained.

But there's a lot of issues that we are glossing over, all of us, about the technical challenges, not only technical, but social and political challenges in doing these kinds of things.

What I would like to talk about are some proposed ways of thinking about this assessment system. In many cases, when you come third or fourth, it is sort of like "What they said."

(Laughter.)

So a lot of what Jeff said and Jim and Laurie, and then Scott and Henry are going to go, "And now for something completely" -- no, not different.

But I would like to talk about some ideas about assessment system. I want to talk specifically about accessible items. I am interested primarily in how we can make the things that we create, both at the summative and the formative level, accessible to all students. I am, obviously, coming at it from an English language learner perspective.

But I believe that the way that we approach the types of items that we think about may actually have washback effects to classroom instruction and what we really are finding out with the assessment instruments. I would just throw out some things to think about. So that is kind of the plan and the direction I want to go.

So, real quick, when we think about a current assessment cycle, what happens is many states, if not most states, test in the

spring sometime. If you have most vendors, that means that you get your results at the end of the school year. So you test sort of midyear, and you get the results after school is out.

So, when the teachers come back for the next year, let's say in grade four, they actually have a new set of grade four students for whom they have no information other than they were in third grade. So there is a disconnect between when the assessment is administered and when the results are actually used.

So, if we want to use the summative information in a way that is meaningful for either the program or for school accountability purposes, et cetera, we are year behind.

Now states, and the State of Wisconsin -- and I was the one who did this -- have said, no, actually, let's do this in the fall. So, instead of doing it in the winter, let's do it in the fall. So we did it in the

fall, and it is my fault.

(Laughter.)

You can blame me. It is my fault.

It turns out, however, when we do it this way, we are actually not testing grade four content; we are actually testing grade three content. So the results that you get, if you get them in December, which is sometimes nice, but often not true, you are getting information about how your students were doing last year, not now.

So the way we have our summative assessment cycle set up seems to be not really conducive to helping the kids and the principals and the administrators in the year that they are actually having the kids that they are assessing.

So several people have suggested that we rethink that and sort of think about a multiple assessment cycle.

Now here I want to mention something. I am constraining my thoughts and my comments to summative assessment. That is

because I don't have time.

But, honestly, in my view, you have to have a summative, formative, and interim assessment system that is integrated with the standards and supportive of the instructional process. That is what you have to have. Any of the grants, I would strongly hope, that would be awarded would have that component as a part of it.

I am working on a grant with the Carnegie Corporation to do just this for English language learners. I think you have to have all pieces.

I want to constrain my focus just to summative assessment because I believe that we are not getting information that is helping kids, all kids, in understanding what they know and what they are able to do.

So the thing you might want to notice real quickly is there's this long time lag between when you give the assessment, whether you give it in the fall or whether you give it in the spring, and a lot of that time

lag is called pick-and-pack, ship, score, figure out the mismatches, and then get the stuff back.

So there is this long cycle between when the boxes get in, get sorted, get out, get tested, get back, get out, and when the actual results come in. And you see in my cute little illustration here I have a shorter timeline, and it is because the presumption -- this is a consortia-based system, so we are presuming that.

Right now, think of this: you can go to Shanghai, take out your wallet, put your bank card into a bank, and get money. You can go to Abu Dhabi and do the same thing. Why can't we deliver an online assessment system that helps students and teachers evaluate their students right now, if we can go find money -- and money is kind of an important thing -- anywhere on the planet? We have the capacity and the technology to do this online right now.

Real quick, I want to say what

Laurie said. We have to have shared core content. We have a common set of content, but in a consortia-based system, you have to agree on what content you want to assess.

I don't know if any of you have been involved in standard-setting or standard development committees. I have. It is a very exciting endeavor.

(Laughter.)

People have very unique opinions about what goes where when, and what's important in the State of Wisconsin, we have a new law passed now about labor history. Now I didn't really know labor history was really important and needed to be taught in so many grades, but it does now.

So getting a group of folks together from different states and agreeing upon shared content is something that, if we were to do this, we need to do. I am just letting you know it is not the easiest thing in the world to do.

And I guarantee dollars to donuts,

if you have a common shared core content, State A, B, and C will want the unique component of their content associated with that. These are things that I think you are going to face when you have a common assessment system, especially if you have a common assessment system that has more than one assessment over the years.

Then, not only that, you have to agree on something that I am sure everyone is really open to, a common scope and sequence. So what's the fall curricula? What's the winter curricula? What's the spring curricula?

Another challenge, something else that needs to be done, and it is very difficult, like Laurie said, there are some things -- math; I'm actually a language arts person. Well, I am a math person, too. I have been working with language arts data and it is not linear. It kind of goes sort of non-linear. It goes really fast and then it slows down.

Some things like algebra, it is kind of like stepped. You need to know how to add, subtract, multiply, and divide before you can figure out A plus B equals C.

So it is like there are different ways of sequencing these things, and they are not always the same. In order to have a common assessment system across states, you really need to sort out what your scope and sequence is. That is a good thing. I think we need to begin doing that. But let's not kid ourselves in believing that that is easy.

The other thing is report structure. Here is something I want to pack on what Jim said and what Jeff said.

I would like to spend a lot of time, and I am not going to -- teachers and administrators have challenges in using the information, enormous challenges. In many cases, I have worked with states, districts, and schools, and I have provided them with assessment information. I have actually kind of outlined that and showed them how they are

doing. And they are like, "How did you find that out?" And I show them. Then the invariable next question is, what do I do next?

We really need to provide, if we are going to do an assessment system that is innovative, that incorporates these common core standards, that incorporates a common scope and sequence, we need to provide the capacity -- what they are doing in Alberta. But I would love to see not 20 percent, but 100 percent of the teachers involved.

I think through the formative assessment, in my view of formative assessment, that is where you get the 100 percent.

I am going to shift to English language learners real quick. I am going to show you -- I am really at the 10,000-foot level and I am going to go to 10 inches in a minute.

The kinds of content assessments we currently provide, in my view, limit

accessibility for English language learners and don't allow us to get at the underlying knowledge that these kids know or don't know.

Steve Elliott, who is researcher in special education, introduced me to this idea that an assessment has an access skill that is the way to get the target skill, what I want to know.

So, in thinking of that, I have sort of thrown out some ideas about developing some types of assessment item types and item formats that are language-neutral, and then those that are supported language and then those that are traditional.

I say that because I believe knowing the academic language content is a part of learning content. But if you want to know if somebody knows how to add, if a child knows how to add, subtract, multiply, and divide, asking them a paragraph-size word problem may not be getting at that.

There's also, when we talk about math and language arts, and I don't want to

discount science and social studies, but in the English language world there's a 500-pound gorilla in the room. Where is the distinction between the English language proficiency and English language arts? We need to begin addressing that issue. We're not.

I mean most of us linguists are chickens, but we really need to begin doing something about that. That is a really critical issue.

So what I want to talk about right now with the remaining time that I have is dealing with accessible item types.

Not quite yet. Not that one yet. Yes, thank you.

All right. What I want to do is to show multiple choice items are okay. They are useful tools. I have used them a lot. I like analyzing them. Can we think beyond that? Let me show you.

I am going to show you a math item. This is a math item I downloaded from a state. It is supposed to be an online math

item. It, in a sense, is perimeter and area.

I am focusing on math in my illustrations that I am going to show you because, as I said before, I am a linguist and I am a chicken. So English language arts and English language proficiency, I am not going to deal with.

But right now let's look at this item. If you are an English language learner, first off, the question is, what is important in this item? What do we really want to know?

It seems to me the thing we want to know, and I am sorry it is really small, the thing we really want to know is perimeter or area, that notion of perimeter or area. But we have got words like "net" and "prism" and "rectangular" and "represented". These kinds of words are in the prompt, and in many cases in the English -- first, actually, the word "prism", if you look at the actual prompt, is irrelevant to the answer.

It doesn't do anything for you. But if you are an English language learner,

you just hit that word. You are going to spend a lot of time trying to figure out how that is relevant to the task at hand, when the task at hand is just finding the area.

So I am suggesting that we can think of approaching items and approaching ways of looking at items from a different perspective. Let's move on.

So I am going to be showing an item from a project that is being conducted with a consortium of the State of Illinois and the Center for Applied Linguistics, taking a look at sort of getting language-neutral kinds of assessments.

This is the same kind of notion. We want to evaluate area and perimeter.

Let's go to another item. Let's go to apples.

That's great. Isn't it great? Technology is going to work.

All right. Well, we are not going to get the opportunity to show you this, unfortunately.

There you go. Stop. Don't move.

(Laughter.)

Okay. The point is, if we want to deal with perimeter, if you could take the icon and go over the word "perimeter"? Okay. Let's identify what we mean. Let's go over "shape". We are talking about that thing right there.

This particular problem, which is the second part of the problem, essentially, is there is a piece that we fly over, which was the training piece. It said, basically, what I want you to do is just sort these shapes out.

Now what they wanted you to do, or this task wants you to do, is to actually identify what the perimeter of that shape that was just created -- this is the second slide -- is.

Now can you press the sound piece right there? And turn the sound up.

Anyway, it actually tells you what is being asked. The idea is, do we care that

we have all this language or do we care that the student can actually take shapes like this --

(Sound from computer is, "What is the perimeter of the shape?")

It sounds better with headphones.

(Laughter.)

But the point is, what are we interested in? Are we interested in the student knowing the perimeter or are we interested in students getting through the language to get to the answer?

So there are lots of items like this that really help us access this idea of, what is it we really want to know? Is this is a multiple choice item?

If you look at the depth of knowledge, I am going to show you another item, if I can.

Can you escape and close? Can we do that? Escape first.

Unfortunately, it is not in a test system, so it is being weird.

Then you close the whole thing out. That would be really cool. Good.

All right. Here's another one. This sample item, basically, it deals with quantities. We want to know whether a student can do a task and then sort out what the quantities are, greater, less, equal, et cetera.

Again, we've got Garret, Sam, who may or may not be common, packages, batteries, et cetera. The point is we want to know if the student can manage quantities and understand the relationship between the quantities.

Then we have another item. Let's show that one. It is somewhere there. It's a thumb drive, it sounds like.

There it is. Kingston. Kingston, there we go. Wow, ONPAR. Apples and oranges. Driver. Bingo. Don't touch anything.

(Laughter.)

Okay. Apples and oranges. Okay, can you press Go?

Now the question is, can you go over each of the words? How many? Okay. So now how many apples -- and bags, you can go over bags and oranges -- cost $12.20? Now we have to solve this problem.

Just to be really quick, it is two and one. So, if you put two and one in that area by -- oh, the icon to the left. The point is we are trying to minimize the notion of language load. If you press that, it will tell you what we want you to do.

So, if you go down and just type "two" on the calculator thing down there, "two", and then go to the other question mark; press "one". Then the check.

Show the price of one apple is less than the price of one orange. Are we concerned about getting through the language of quantities or are we concerned about less than or greater than?

So you could actually click on the apple to the left. Is less than, the "less than" sign, and the orange. Or you could put

120 is less than 125. There are variety of ways of answering this item correctly.

Again, is this a multiple choice item? What's the point? What are we trying to measure?

The idea is that I believe we are at a place right now -- and you can escape and get out of that, if you wish. This is a project, again, that is being spearheaded by Rebecca Kopriva at the Wisconsin Center for Education Research and the State of Illinois.

The notion is, can we get at this language at that level of the content that we are really interested in and not dealing with the language load that we often have to get through, the access skill that we often have to get through to access that information?

And we will get to my slide in just a second. That is Laurie's, I think. Later, later, later. Further down.

I have five seconds. Yay!

(Laughter.)

Okay, move. There we go.

So things to think about.

Back up one. One more. Stop. There. Thank you.

(Laughter.)

Oh, well, 26 seconds over.

MS. WEISS: Don't worry. You get extra time.

MR. COOK: Thank you.

I think we need to think beyond multiple choice. I think we need to think about not only English language learners, but students with disabilities. What do they know? Is the format that we are trying to get that information prohibiting us from getting the information that we want to know?

The technology to do this is available now. This is being done in one state. It has been done in science in another state.

Current measurement models will work. It is not easy. I don't think anything we are talking about is going to be, oh, that is simple; we have already done that.

How do we create inclusive assessment systems that assess the broadest range of students possible, so we can get information about those kids and help them?

I am going to end on this. Students, parents, teachers, administrators, we need to expand their capacity. Because we could create the coolest-looking, powerful, adaptive, you name a cool click word, and get a number and provide that to teachers and administrators, and they don't know what to do with it.

We need to have a system that integrates student professional development, capacity-building, institutes of higher education integration, and parents.

So I'm done.

MS. WEISS: Thank you. And thank you for running the computer back there, putting you on the spot. That was great. We all got the hang of what Gary was trying to say.

So let me see what questions we

have got.

MS. WURTZEL: So I have a question, Gary, that I actually wrote down before the technical difficulties started.

MR. COOK: Okay.

MS. WURTZEL: Which is you emphasized, and other experts before you emphasized, the desirability of moving to online assessment. So my question is, given your work in Wisconsin and the other states that you are working with, what is your sense of the capacity to move in that direction, and how quickly that could be ramped up, so that you could see this at scale?

MR. COOK: We are kind of there. I think we are close to there. I know some states are still behind others, but I think we are close to being there.

The T1 lines, the lines going through, the big issue is the pipeline and the hardware at the schools. That is the big issue.

But, right now, it seems that more

and more schools -- actually, it seems that rural schools are probably more there than urban schools. But I think we are close to there.

MS. WHALEN: Could you talk a little bit more about helping parents to use assessment information more effectively?

MR. COOK: There are a couple of reporting systems that I have found out about, like the Grow Network and other places, that actually have things like, let's say you get a result back from a language arts assessment and reading. You know, the technology is available now where you can say, okay, where's my child at; what kind of books at the local library could help them read what they are interested in?

I think the idea is we need to incorporate the parents in determining what the information means for their children. Often, they are not.

Also, they need to be able to get the information in a way that makes meaningful

sense to them. Unfortunately, because I am a psychometrician, I like throwing out numbers and Greek symbols, but not many people do numbers and Greek symbols. They do different things, like, okay, if you go to the local library and you look at this book right here, your daughter might really be able to get that. I think we need to begin doing those kinds of things.

MS. WEISS: Other questions?

(No response.)

So let me ask you, Gary, you were talking about the implications for ELL and students with disabilities. Have you done much thinking about assessing -- and I don't mean in a high-stakes way -- but assessing kids at kindergarten through third grade? Because it seems like a lot of what you are doing also could apply down there, and whether you have got thoughts or advice on any of that?

MR. COOK: It's hard. Don't aggregate it in any kind of accountability.

MS. WEISS: Right.

MR. COOK: Yes, sure. The challenge with kindergarten, first, and second grade in a school setting is we have several things going on at the same time. We have the cognitive development; we have academic development; we have social development, all happening at the same time and all happening in a dynamic way.

Let me give you a great illustration from my child language acquisition teacher. She studied three children from 12 months to five years. Every month she went and interviewed them in a long corpus. Then she aggregated that. This was before they had computers. She did this all by hand.

What she wanted to do was she did things called types and tokens and ideation. How much stuff a kid was saying and what kinds of vocabulary they had when, basically, they started talking to when they were close to fluent.

What she found out is that all three kids started at the same place at 12 months and ended at the same place at five years, but they didn't get there the same way. That is the challenge with K, one, and two, is in many cases many kids don't start at the same place, and they have different pathways. So you need to provide a system and a mechanism to try and help teachers and parents sort out how to best help their kids where they are at, both developmentally and academically, to move to the next step. That is just a hard thing to do.

Sorry, that was depressing.

MS. WEISS: No, no, no. That is good. That was great.

Other questions?

(No response.)

We have worn ourselves out.

Okay, it is time for lunch. We will reconvene in this room at 1:15 and get started at 1:15 sharp. So you might want to come back in about five minutes before then.

Thank you.

(Whereupon, the foregoing matter went off the record at 12:08 p.m. for lunch and went back on the record at 1:18 p.m.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

1:18 p.m.

MS. WEISS: I take it as a good sign that you all came back from lunch.

So our next speaker is going to be Scott Marion. I am going to turn it over to Scott, let him introduce himself, and we will get going again.

MR. MARION: Hi there. Scott Marion. I am from the Center for Assessment, just, actually, up the road in Dover, New Hampshire. I am pleased to take a bus to a meeting for once.

(Laughter.)

So, actually, I couldn't help it this morning as I was getting ready to leave, the first page of today's business section of The New York Times, I just had to laugh when I read this, David Pogue's technology column. He just says, "I mean, come on. We're not asking for the world, just a handful of prerequisites for the perfect camera." Then he lists about 20 things, starting from small

enough for pants pockets, big sensor, interchangeable lenses, and he goes through about 20 of these things and says, "Is that so much to ask? All implausible feature-lust aside, though the main obstacle to obtaining all of this in a single camera is this little nuisance called physics."

(Laughter.)

It reminded me of some of the problems that we are facing today, as we are structuring this, trying to help the Department structure this notice.

Like Laurie and many of us who have some background in cognitive psych, a little advance organizer, talk about a theory of action all the way through some, hopefully, practical advice on the proposed -- I have been calling it RFP, RFA, but I am not sure if it is a notice, or what it is.

MS. WEISS: It is an NIA.

MR. MARION: NIA, okay. Good. We will have to change that.

MS. WEISS: Notice Inviting

Applications.

MR. MARION: There we go. Okay.

So I have the privilege of working with Rich Hill. Rich likes to say, you know, sometimes you've just got to put a stake in the ground and then let people react to it. So I am going to try to put a stake in the ground here.

So a little preview of my vision; I am going to get into much more details later. But I would argue that it is conceptually coherent. You will have to be the judge of that.

It is a comprehensive system that includes these end-of-year summative assessments we have been talking about. Interim performance tasks, but, notably, embedded in what I call mini-curricular units. We might have to change that name. I learned it, opportunity-to-learn units. Formative assessment supports or probes, and then, obviously, professional development and actionable reporting.

Those of you, a lot of folks I work with here are probably sick of me talking about things like theory of action. Laurie is shaking his head yes.

But I would argue that we need to have, as part of this notice, a requirement that states a consortia -- come up with an explicit theory of action. I would actually challenge the Department to do the same, is describe what are these very clear goals. We have to specify what these goals are, and we will come back to that at the end.

Then how do we think that these goals can be achieved? What are the specific mechanisms? How are we going to get from A to B? Just clicking our heels and hoping we get to Kansas or Boston is not necessarily the best way to do that.

So that will help us sort of explicitly describe particular prioritized design choices. So whether the goal here is to influence and shape teaching and learning or measure existing knowledge, those two can

be competing, or making cross-state comparisons, they don't all work together. But having a coherent theory of action is a check on the logic of the underlying assumptions and on the requirements.

And like Gary said, we will be able to say, "Yes, what he said." In this case, I am surrounded by two experts in the field or, actually, multiple experts.

But a lot of us have been talking about purposes and uses. I am not going to go into a lot of detail here. I am going to put a little pressure on the Department folks. I know that Congress will reauthorize ESEA, and it is their job.

To the extent that we could foreshadow the accountability uses and be as clear as possible before we let this NIA, it will help people as they try to put forth their proposals. Because it is going to have accountability uses, and that is going to shape a lot about design and validity issues.

The other thing about being very

clear about the purposes and uses is there's lots of design deliberations that are going to be in effect as we go forward. If you don't have a clear sense of your purposes and uses, you don't have a touchstone by which to help make those decisions.

So, again, putting a stake in the ground, I am going to say that my overarching goal, and we had lots to choose from in The Federal Register notice, is that all students should have meaningful opportunities to develop deep understanding of important content and critical skills to allow for viable post-secondary choices -- if you want to say college and work-ready, go ahead -- and for becoming contributing members of society. So I am going to propose a system that is intended to support this overall goal.

So I actually said, all right, what are my prioritized purposes? And I would expect states and the U.S. Department to go through similar exercises. There's many plausible purposes here.

So I say I want to measure a limited number of big ideas, right, in order to help students use this more robust knowledge in novel and complex settings. And sort of a subpart of that is asking for better integration of curriculum instruction and assessment.

Then measuring student longitudinal growth as a foundation for valid accountability systems and as information for school improvement.

All right, so I am limiting myself to two main purposes. Maybe I cheated and threw a third one in there. But you will notice that I am intentionally not focusing on cross-state comparisons. A lot of people would argue, well, that's why we have the common standards, so we can do these cross-state comparisons.

I think if you go back to my overarching goal, I don't necessarily agree that cross-state comparisons are as important as other folks make them out to be. Now that

is my stake in the ground. You could say, no, I think cross-state comparisons -- and I am going to show you where they fit in my theory of action to make learning better.

I would argue that the design principles rest on some of the best information that we have currently about how to design assessments building on important theoretical foundations. So I threw out two examples. Luckily, I got a preview of Henry's, so I am not going to have to go into as much detail on evidence-centered design. Henry has nicer pictures than I do.

The CREST work on certain ontologies is related to this kind of work. "Knowing What Students Know," the NRC publication, Jim Pellegrino and Naomi Chudowsky, and Bob Glaser. Any one of these would be a more substantial foundation than how current assessments are currently designed.

One of the things that is going to be critical as this NIA is written is that, if

the Department would require that assessments be designed according to these principles, and I wouldn't specify that it has to be designed according to evidence in a design, but it has to be a sound, theoretically-based design, that people aren't able to just get away and say, "Oh, yeah, we do that." It has to be much more specific about how you intend to incorporate these design principles and how they are going to help make your assessment more coherent and better.

So my proposal is to build this coherent system that bridges curriculum, multiple forms of assessment, and supports for instruction, and includes these pieces that I laid out before. I will go sort of piece-by-piece here.

So, first of all, as part of my theory of action, I argue that you have to think about the reports upfront. Ron Hamilton has a great line. He says, the only way we really have of communicating with the public is through our reporting system, and it is the

last thing we design. It really needs to be one of the first things we design.

I would argue that it needs to be conceived as a system of reports. There's different purposes, different users, the different levels of information. And they must be actionable -- you have heard some talk about this already -- at least through appropriate inferences, decisions, and instructional or programmatic actions.

This is a website off of the Colorado Department of Education's website, schoolview.org.

A little self-serving here, Damien Betebenner from my office was one of the key designers of this reporting system. Many folks have seen it. Jeff has just started using it in Massachusetts or something similar.

We argue that reports don't have to be the static, one-page piece of paper. The reports should support the theory of action.

Sorry if this is a little small. I

am going fast, you know.

This is the part that I would say is probably the most radical of what I am proposing. I will talk about this a little bit when I talk about access and equity more. But I am suggesting that, for those of you who are old enough to remember College Board's old Pacesetter Program, it is sort of akin to this, where you have two -- and I am just throwing out numbers here; this could be debated and it could be phased in, but it has to be enough to make it worthwhile and not too many that you have taken over the whole curriculum.

But these could be short curricula units, be as short as a few days, as long as a couple of weeks. Each unit must focus on a big idea of the domain, and not just trivial matter.

They have to be able to be strategically used with existing curricula, so you are not writing curriculum, used with existing curricula.

But, importantly, they serve as the basis for performance tasks and also a context for summative assessment. Too often we are so worried that we don't know how people have come to the knowledge that is going to be tested on the summative assessment. We make it so general, and any sort of context we put on the assessments -- so you are asking people to sort of reason with evidence, which is sort of like an IQ test. That is why we don't see a lot of curriculum and instructional sensitivity on these summative assessments. This could help provide some of the context for that.

It would include training materials and supports for implementing formative assessment or progress monitoring. And it would be flexible enough and robust enough to use multiple years, but changing out the assessments.

Importantly, and I will talk about this a little bit more later, it provides a vehicle for structuring opportunity to learn

and access for all students, because now you have some control over what they are being exposed to.

Lots of folks have talked about the summative assessment. I don't think that my proposal is particularly different than some of those smart ideas we heard already.

I will actually just focus on this one thing with the summative piece. I keep hearing this. You know, we have this obsession with instant results. We have got to get the results back right away.

When I was a test director, people would say, "Well, I don't get the results back until the end of the year. I don't know how the kids are doing." We tested at the end of the year. Of course, you don't get the results back until the end of the year. And if you are waiting for me to tell you how the kids are doing, I wonder what you have been doing the first nine months of the school year.

So I think that we could have these

results -- I shouldn't be putting them on a NAEP timeframe. No offense.

(Laughter.)

Well, NAEP has different purposes.

(Laughter.)

But they don't have to be coming back right away, which would allow us to incorporate things like open-ended items and actually move through the quality control process. Actually, now the other thing is we can turn these results around very quickly. A lot of vendors in the room, they could do this really fast, but nothing is free. So it is going to cost the same. Nothing is free.

These interim performance tasks, I would argue, are the foundation of the system. This is where I differ from Jeff a little bit. I would argue that, first of all, they are contextualized within the curricula units. They should be scored locally and incorporated into local assessment and accountability systems. For instance, in Wyoming and Rhode Island, they are using these

sorts of things in graduation systems.

As Laurie talked about, they could be scored locally but audited, if you want to include them in the state accountability system. Rich Hill and Brian Gunn talk about in the Kentucky portfolio system, this is one of the fastest ways to -- Sue Rigley as well is sitting over there quietly -- as one of the best ways to bring about change and developing a shared understanding of standards and criteria, was to actually have this local auditing, essential auditing of this local scoring.

Obviously, they should be designed according to some of the principles that I have articulated. Most tests should be released each year for use in the schools.

Formative assessment, here Laurie and I disagree a little bit, at least in the language. I would argue it really should be formative instruction here. We are really trying to facilitate these probes and processes, so that teachers, actually, can do

this minute-by-minute work or day-by-day work with students to improve what they are doing.

Here I agree with Jeff; there has to be a clear separation between formative assessment and any kind of district and state accountability systems.

So I was asked, I got a call from Mark Sternberg at the Department yesterday who said, "Remember, by the way, you're the assessments-with-disabilities expert on the panel. So make sure you talk about that." I have been working on this thing for a little while.

But I would actually argue that we have much less of an assessment problem than we have of an instruction problem. I have been arguing that we don't need a 2 percent assessment; we need a 2 percent instruction.

I think we have been focusing on asking assessments to make up for lack of opportunity to learn. They just can't do it.

So I am suggesting that these proposed opportunity-to-learn units -- I have

changed the name -- are designed to help level the sort of curriculum instruction playing field, should provide supports to teachers to help ensure that they are able to provide supports to their kids and the formative assessment information that they need to be able to diagnose that. Allow us to design tasks with multiple and varied opportunities for students to viably participate in the system. I think these opportunity-to-learn units would serve that vehicle.

Is it going to be easy? No, but I think, if we want to spend $350 million, I would argue that we need to be thinking about the instruction and the curriculum side as well.

Yes, it is new psychometrics. It is not really that new, actually. Because I was laughing; I was in a meeting this summer, and it was talking about the new psychometrics ways of thinking about this. I was sitting next to Eva Baker, who I know is speaking tomorrow. Well, she and Bob Linn and Steve

Dunbar wrote an important article in 1991. Sadly, I actually remember the color of the journal when it came out.

(Laughter.)

So we have been thinking about this for a while. We tend to privilege reliability or pretty scales and overly-strict notions of comparability, and we have traded some important validity considerations for that. We have some tools to be thinking about this differently.

I know you are going to be talking about high schools tomorrow, but I want to say that one of the things I think is really important is, for now, we have been treating high schools in many cases under NCLB as just sort of grown-up elementary schools. I am arguing that they are different. The assessment system needs to be considered differently.

Again, I am putting my stake in the ground. I think that the assessment system should be situated as a specific indicator of

core courses up to some point. I know that I differ from the common core folks; I know I differ from the Achieve folks, friends of mine.

But I am saying let's make this common up to 10th grade, or some number that is not 12th grade. After this point, there should be more choice in the assessment accountability system to allow for sort of high levels -- nobody gets a low level -- high levels, but for specialization.

I am a math and science guy married to an artist. I understand the importance of diversity of expertise.

(Laughter.)

These interim performance tasks can be used as part of the student accountability system at the high school level.

Marion Snyder from Rhode Island is here. They are doing something similar. Wyoming has been doing it. It can be done.

It is free advice because you are not paying me.

(Laughter.)

So it is free advice that recognize that development is an ongoing cost. It is not a one-time purchase. No matter how much we put into this, it is going to be ongoing.

Recognize and embrace these differences that I just talked about between high schools and elementary schools.

Figure out what are the absolutely essential pieces that need to be in this new vision of the assessment system, and then you have to examine the cost for every additional component as opposed to just try to spread it thin among all, and not do any one well.

Reconsider the current practice of having every student tested on every item. Now there's certain cases where that is important, but there's lots of us in the room who still think matrix sampling is viable for certain things.

As Laurie said, allow for multiple awards because nobody has the right answer. If they tell you they do, they definitely

don't, and especially in the right context.

(Laughter.)

This is especially true for high school because we don't really know, should it be course-based, should it be common in high school?

Again invoking Rich Hill, he has responded to a lot of RFPs, and here's, importantly, to be exceptionally clear on the goals of what you want and be as flexible as you can on the specific means to achieve those goals, unless you are absolutely clear on what you want, and then don't play games with it. Tell people what you need to do.

Think about a phase-in over the next five years. It is going to be hard to do this any faster than that.

Then recognize these critical operational, bureaucratic -- bureaucratic in a nice way -- constraints. We have existing contracts. We have certain state laws. And if you are like me, you deal with state procurement rules. It is a whole other

nightmare. So there's lots of operational things to keep in mind.

And that is it. Write things formally, but if you want to reach me, that is how you do it.

MS. WEISS: Thank you.

Questions?

MS. JONES: When would matrix sampling be appropriate in your design?

MR. MARION: Oh, I could argue that for lots of parts of the summative assessment. Particularly, I could see using matrix sampling to get more school-level information. I could argue that not every kid needs to take the same interim test, that within these particular units I might have different tasks. You know, if it was a science unit and I was really more interested in population biology, and you were more interested in some other aspect of biology, I could see varying the test that way.

It is like we ask kids to write a research paper. Not every kid writes a

research paper about the same exact topic, right, usually in most classes. You say, what is the topic you want to write? We are going to do a term paper. What is the topic you are going to write it on? It could be different by different kids. We say we have common ways of scoring that. So I would argue that we can think about it there.

I am not saying we have to do major sampling. I am just saying we have so far taken it off the table, that it has a certain efficiency, and I don't want to see us lose it.

MS. WURTZEL: Scott, at the beginning when you set up purposes of assessment and you put your stakes in the ground and put cross-states comparison as a lower priority, can you say, given the assessment system design that you have laid out, what would have to change or what would restrict it if you also wanted to get cross-state comparability?

MR. MARION: Well, if it is a

consortium, and this summative assessment is common, at least according to some, you are going to have it easily.

MS. WURTZEL: Right, right.

MR. MARION: But it is not something that I would build in as the most important of it. Because there, if that is really my most important thing, I really want highly-reliable scale scores that I could compare, and I am willing to give up some of that reliability if I think I am going to get measures that are more robust to understand that.

Then I would ask Henry and Laurie just to do it, and they would figure out how to do it.

(Laughter.)

MS. WEISS: So that leads right into the question that I had, which is, on one of your slides, you say that we should push for requirements and expectations beyond the current safe methods of psychometrics. What does that mean?

MR. MARION: Well, a lot of my dear psychometrician friends and brethren would really focus on the importance of reliability or minimizing, equating, and things like that. Those are all important considerations; there's no question about it.

But lots of us know that, if you really want to maximize reliability, besides having an unbearably long test, you make things more homogenous and you lose certain aspects of the diversity of the domain.

Now Mark Reckase and his folks who are doing multidimensional scaling would say, "We have ways of doing, of accomplishing both." Not done operationally yet.

So I think that if we are willing to loosen this focus on sort of this unit dimensional domain, it has pushed us in certain corners that we are having a little trouble escaping from, and it is not clear to me that we need to be in those corners.

MS. WEISS: So what would you heighten?

MR. MARION: I would stress the importance of utility, validity, real construct validity of these important things, the kinds of predictive evidence that -- I think it was you, or I saw it in Henry's presentation -- predictive evidence of, do kids who succeed in this way, are they more likely to succeed at the end in this college readiness?

Or in the way that Laurie talked about, which I thought was fascinating, thinking about this kind of scaling in terms of actual performances and evidence related to the knowledge and skills as opposed to a numerical scale. To do what Laurie put out there -- so you didn't ask him all these questions (laughter) -- but to do what Laurie put out there --

MS. WEISS: I think he refused to answer.

(Laughter.)

MR. MARION: That's right.

We don't honestly have the tools to

do that yet, but I think that there's enough smart folks around that we could actually figure out ways to do the kind of thing that Laurie is talking about and in some ways similar to what I am talking about, is not just say, "Oh, but it's not linear." We have to think about it.

MS. WHALEN: A slightly different question: in your theory of action, are the curricular units that you see being adopted to measure these domains, are those common across all states and districts involved in the consortia? So you would be saying that you have to do this at this time?

MR. MARION: I would say you have to do this or, if you participate, you should do this. And actually, I have been debating that.

So I wouldn't have varied units. But I would say, "You have to administer these" -- I am making up a number -- "four performance tasks. If you don't want to use this unit, that's fine. You still have to

administer these tasks. But we are telling you they are situated within those units. If you don't want to, that is fine."

Similarly, I am not saying, on October 1st, you are going to start unit one, and everybody else in the state or the consortia is going to do that. You embed them when you think they fit most.

Again, depending on how you count this in the accountability system, it is going to raise issues. People are going to worry about security and things like that, which is why I think that we are smart enough about accountability that I think we have figured out ways to ameliorate some of those issues.

So I would argue that I wouldn't, for the cost and the resources, I wouldn't be developing different units for every district. Then you sort of lose the whole point of it.

The idea about it as well is it allows a certain efficiency in professional development. The only way that we are going to get at real formative assessment is with

some efficiencies around professional development that are situated within at least common aspects of the curricula.

MS. WEISS: Jim, did you have a question?

MR. DUECK: Right. This conference is about the assessment initiative, but curriculum gets into the picture as well. When curriculum lacks specificity, the teachers are often driven by the assessment to find out what was really intended in the curriculum.

Is that a concern or is it just simply a perspective?

MR. MARION: It is absolutely a concern, which is why I said here's the stuff and don't guess; we are telling you what it is.

And again, they have to be really important model units based on big ideas and not trivial kinds of things, that people will then model and build other units.

MS. WURTZEL: So, Scott, you talked

about the importance of reporting formats and building the system with the reporting in mind. Do you have some criteria or sort of core items in a report that we should be thinking about?

MR. MARION: Yes, and I credit a lot of this to my colleague Damien Betebenner, who has taught me a lot about these issues of reporting. We have been working a lot on this together in Colorado and now New Hampshire and some in Massachusetts.

But the idea that, if you, as the experts, think you know the important story to tell, we would like to think -- and I am the assessment director who once bought SPSS for every district in the State. That was the dumbest idea I have ever did.

(Laughter.)

And people could slice and dice their own data. What happens is you get a lot of diced data, and nothing ever gets put back together.

So we think that a report should

actually -- what are the compelling stories to tell with the data to tell it in visually- and graphically-compelling ways that almost anybody could understand with a little work? And to present it from the individual student level up to the highest levels of aggregations you care about and disaggregations. But that it is so graphically- and visually-compelling and accurate, tells accurate stories, that I think that that is some of the design principles. We talk lots about that.

MS. WEISS: So I think we should move on, because we started a couple of minutes late, and turn to Henry Braun.

MR. BRAUN: Well, I am Henry Braun from Boston College, and I am hoping we can find the beginning of my talk. It is somewhere there.

MS. WEISS: The last one, the bottom one. Yes. Very good.

MR. BRAUN: There we go.

MS. WEISS: Thank you.

MR. BRAUN: Wonderful. Thank you

very much.

Well, you can see from the title that I may be one of the more conservative here.

(Laughter.)

MS. WEISS: I see you have my job.

(Laughter.)

MR. BRAUN: But you will see that I am not as pessimistic as you might think.

I first want to thank you for the opportunity to speak. I think the design framework that the Department has put forward constitutes a really worthwhile goal. I think, as other speakers have indicated, we really have quite a lot of knowledge about how to build an idealized system.

I would argue that, if we were starting with a blank slate, we could do it in three to five years. If you just look at the NRC reports, and I particularly like the one called, "Systems for State Science Assessment", that really lays out a lot of what we have been talking about in real terms.

So, for people who say, "Well, you know, this is all pie in the sky. We don't know how to do it.", I would say, no, I think we really know how to do it. We haven't necessarily done it, but we know how.

And I think, if we look across the states, there are islands of excellence and innovation. If we look at other countries, God forbid, there are islands of excellence and innovation.

(Laughter.)

But I think, practically speaking, realistically speaking, in this country we are going to have obstacles to getting there in three to four years because of technical/logistical obstacles, real capacity constraints, contractual issues, which we have talked about, resistance, and even inertia.

So my argument is, okay, let's aim for the sky, but let's think about what the first step is, right? There's a ladder. What should the first rung be?

I would argue that what we want to

do is to chart what I would call a new developmental pathway that will take us off the path that we are in now, which tends to be very incremental and very conservative, and jump-start the innovation and lead towards a good approximation to the ideal.

To do that, I think we really have to think in a systemic and systems way, think about all the interacting systems that impinge on assessment and that, in turn, assessment impinges upon.

If you just think about what is going on now about trying to develop a new system to support all electric cars, and what needs to be in place, and all the pieces have to be in place, you know. Otherwise, it isn't going to work. They don't all have to be at the highest level of utility or cost efficiency, but they all have to be in place in some way in order for the system to take off. That is what I think we need to think about here.

So I think we want a system in our

first rung, I would say for the consortium, that will model new patterns of collaboration, which I think is extremely important; that will have superior measurement properties, will illustrate and employ new paradigms for assessment, and exhibit the potential of these new technology platforms. I will say a little bit more about platforms in a minute.

But before we begin with assessment, we should, I think, recognize that there are some real prerequisites, that we need a comprehensive model of the domain. That is partly because of the instructional specificity that we are going to be looking for.

We need models of student learning, the pathways to expertise, learning trajectories. There are all sorts of names out there.

We need content standards that are not only complete, but vertically-articulated. Other speakers have said this before me; the importance of vertical articulation cannot be

overemphasized.

Then performance standards that are rigorous -- and I will say a little bit more about that -- and also vertically articulated. Too many states right now actually make sure that their standards groups are separated by grade and never meet with one another, which seems to me totally ludicrous. And that we have a technology platform that will support both instruction and assessment.

So this is just a picture that I stole from somewhere, but what it meant to illustrate is the notion of what some call horizontal coherence. That is, you have a basic set of content standards from which flows curriculum and instruction, and then you get your performance standards; you get your assessment system, and out of that flow, reporting, staff development, and so on.

So what you really want is to make sure that everybody is on the same page because there is a common basis for understanding, no matter what the different

aspects of the system happen to be.

So here is my stake in the ground. I call it a four-component system. So there will be a diagnostic, you could call it formative component, that would provide instructional support. It is going to be frequent and maybe technology-based.

Extended projects. I think it is important that students have an opportunity to work on extended projects that go across two or three classes at least, that are targeted at some of the higher-order standards, that are integrated, perhaps even across subjects, and that are technology-based.

I believe that we are on the verge of having a technology platform that would allow summative on-demand testing, but I don't think all states are there. But I think all states should be somewhere along that trajectory, and maybe by introducing it in the less demanding environment of extended projects, rather than we have to have 10,000 computers on one day, all working in concert;

that may be too much for some states.

I think I like also the idea of what I call on-demand, based on previewed materials. That is, say the state department sends out a number of booklets that describe, let's say, mathematics, different problems, different issues, and that on a particular day you sit down and you get two of those issues, and you have got to write an answer or produce one on the computer. So that one has an opportunity to prepare properly, but there's also the notion that you are going to prepare broadly because you are going to get not one, not just one or not all, but only a few of them.

And then, finally, maybe a more standard on-demand that would be forced choice and short answer, which would allow for rapid turnaround time, as long as we are still in the paper-and-pencil mode. If the technology platform extends in its capabilities, we are in a position to have more rapid turnaround for everything.

So what would we get out of this? Hopefully, we would enhance our construct validity because we would be able to better represent the constructs that are embedded in the content standards, and we would have multiple assessment modes, so that everybody has an opportunity to show what they can do.

I think we would also have improve systemic validity. That is, if we think about the washback of the assessment system on curriculum and instruction, that there would be less of an incentive to narrow the curriculum and reduce the value of what we would call inappropriate test preparation.

I think we could also have much better links to professional development because we would have teachers focusing on student work products. The idea of having moderated markings in collaborative sessions is a way of jump-starting new approaches to professional development, but in some areas they are called professional learning communities.

Then we would be, also, in a position to ramp up toward a full-service technology platform. By technology platform, I mean sort of a generic set of components that could be harnessed with relatively little additional investment for a number of different purposes, whether it is differing kinds of assessment for different subjects or professional development for instruction, or what have you.

So I want to focus a little bit on some specifics, just so I kind of differentiate myself from the rest of the crowd, if that is possible.

(Laughter.)

I focus on standard-setting partly because I am interested in it, but partly because I think it is really, really important. If we don't get standard-setting right, then I think it is going to be very hard to get the rest of the project really going.

What I am arguing for, simply, is

to recognize that, when we talk about performance, proficiency, and grade in, that that is not an end in itself. What we are really looking forward to is proficiency and, in the case of this notice, readiness for whatever comes beyond high school, that you are going to get to the high school diploma and be ready for whatever comes afterwards.

So what you would like is a set of milestones, standards, that say, not only have you mastered the material now, but you are really in a position, given reasonable instruction and reasonable effort on the student's part, to obtain mastery, proficiency, whatever you want to call it, for the next level.

That is why I would argue very strongly for this cross-grade coherence in standard-setting, which I think you could develop, and I have argued elsewhere, sort of through a backward deduction method. So that, if we, in fact, have readiness standards for college, the world of work, the military, and

so on, then we could work backwards from those to say, okay, if that is our goal, what does it mean for someone in the 10th grade to achieve over a broad range of competencies the readiness to succeed when we get to the 12th grade diploma requirements and beyond, and then sort of work backwards.

One of the problems I think we have now is that standard-setting at the state level is, what I would say, it is a hermetically-sealed system. It is self-referential and, in fact, glories in the fact that it has no connection to the outside world, which is great as long as the student stays in the school system, but when they get to the real world, having been told that they were proficient in everything, and find out that when they want to go to community college, they are actually not ready and they have to spend a year doing the so-called developmental work.

So what I have argued for is what we call the 3P paradigm. That is that, when

we develop these standards, they should be prospective. That is, the standards that you hope to measure should be an integral part of the way you develop the tests. Right now, we develop the tests and, in fact, hope that they are a reasonable support, from a measurement point of view, for the standards we are going to set sometime later. That seems to me to be really hoping against hope.

So why not bring your standards in a particular form into the test development process, and do it in a progressive way, so that, in fact, you have this coherence, this linkage or reticulation, whatever term you want, from one grade to the next. So that, in fact, that through empirical work, you can argue for the predictive value of meeting a particular standard in this grade or at a particular milestone, because it does say you are on track to reach the next milestone and the next, and so on, until we get to the end of high school.

Oh, this is just a picture to

represent that, in fact, you can lay out -- this was done in the context of evidence-centered design -- you can lay out a process by which you can think about how to embed preliminary performance standards in a test development process and move through. I won't, obviously, go through it here. I can give you the reference at some other point.

I want to say a little bit about technology. Technology, of course, is not an end in itself. It is a means to an end. One of the questions we have to ask is, to what end? I think there's also a lot of discussion about technology platforms. I am sure we each mean something different, but I think that it is important in the notice to be very clear on what you want consortia to be able to say, and what they are going to develop in the context of a technology platform.

The issues of capacity, or lack of same, at all levels, I think is a real issue. Particularly, if we are talking about a consortium of states, they are certainly going

to vary in their capacity, not to mention there will be, obviously, variation capacity within a state. That needs to be, I think, front and center in terms of how consortium proposes to move forward as a consortium in terms of developing a technology platform.

Speaking very narrowly, the promise of technology to the end of better measurement is that it will allow us to introduce lots of novel item types, not for the sake of novelty alone, of course, but to improve construct validity.

It will also allow us to introduce adaptive test designs, which we have heard about, which can, at the least, improve precision, but also in the diagnostic or formative setting can provide much more useful information at the teacher level.

We also, once we introduce this kind of technology, can now draw on a lot of work on expert systems that have been developed to grade very complex kinds of assessments in mathematics, in graphics, in

natural language, computer science, et cetera, et cetera.

So that the cost issue becomes much less problematic once you have gotten onto a technology platform, and that needs to be laid out as well.

And finally, it can improve accessibility for some because you can have a large fund you are going to employ. So there are many accessibility issues that are ameliorated with the introduction of technology.

Now what are the implications for assessment design? Well, it makes it much more complicated, frankly, but, hopefully, much more rewarding.

I think one of the reasons that it will be more complicated is that, if you buy the argument that we can't get from A to B in one step, but we need to take steps A1, A2, A3, and so on, then what we are going to be asking the consortium to do is to think about a developmental pathway, a multi-phased

sequence, if you will, that will incorporate both evolutionary and revolutionary elements. We are not used to doing that. We are used to saying, okay, we want to do this, and this is how it is going to happen. But here we really need to have a longer-term plan. It is not something we really have a lot of experience doing.

Another issue I think that makes accessible design much more complicated is this issue of integrating cognitive and development perspectives while paying attention to, perhaps not undo attention, to traditional psychometric and logistic requirements.

So there is going to greater complexity balancing goals and constraints, and it means that we are going to need much richer, multidisciplinary teams to be able to handle this particular set of issues.

This picture is just kind of really a schematic about how you might think about assessment design. I call it my 3C plan,

where you really focus on constructs, communication, and constraints.

The thing about communication is what it is you want to be able to tell people as a result of the assessment. If you combine this communication, what you want to say, with the constructs, then you can apply evidence-centered design, as an example, in order to generate particular kinds of task types or assessment designs.

Constraints, for example, universal accessibility also plays out. So you can think about universal design approaches that, together with the constructs, would allow you to think about these different issues.

The point is that, rather than making incremental changes, you want to be able to generate a number of different elements in a design space that are much broader than what we are doing now, and then be able to test those models.

In a very real world, I think that speaks to having at least two, but maybe more

consortia, taking different approaches, so that we really see the pros and cons of, hopefully, radically-different approaches to innovation for the purpose of learning and instruction. Then, of course, another part is evaluation. So you really want to have, build in a very strong audit and review mechanism.

Now I know we are not talking about accountability, but, of course, it is the 500-pound gorilla, or maybe the 800-pound gorilla. We believe, I think there's rumors that the reauthorization of the ESEA is in progress again.

It is reasonable to expect that whatever comes out of Congress will include both status indicators and growth, and maybe value-added indicators. I believe that superior test design, the kind of improvements that we are talking about and hoping for, can enhance the validity of the process, of whatever accountability system comes down the line.

But it is important to remember

that test design alone isn't going to guarantee the consequential validity in an accountability system. Consequential validity of a system really depends on all the components acting in concert. Without good testing, we are not going to get good accountability. With good testing, we have a shot at it, but it is not a guarantee. There are lots of fundamental problems. Of course, the fundamental problem, from my point of view, is trying to causal inferences from the very messy observational study called American schools.

So what are the challenges? Well, let's put it on the table. There is going to be resistance and inertia. Wrenching change is very difficult, and we have the constraints of current contractual arrangements. Just think about a consortium of eight states, each of which has its own contract with a different vendor on a different schedule with different things. How are they going to think about putting that together?

So one of the issues is, how do you identify a viable state consortium or what are the requirements you are going to place, so that they can make the case that, yes, we are viable, despite our differences?

I think it is important for this notice to encourage innovation without being overly prescriptive. So I would not hazard a guess as to say how many item types should we have, what should be the percentage of items released. I think that has to be played out in the process. I think you can say that you want to have multiple formats. I think you can say that there needs to be a release mechanism, but I don't know that you want to be more prescriptive at this point.

I think you also have to recognize that there are going to be complaints. As we expand, with all good intentions, the quality and range of assessment, we are going to get this complaint that the assessment tail is not just wagging the dog, but it is sort of whirling around and sending it off into

another orbit.

(Laughter.)

On the other hand, the complaint will be that, "Oh, you're bringing the teachers into it? How can we trust them? You know, they are going to grade things so that their students do well or their friends' students do well," and so on.

So you are going to get it from both sides, which maybe at the federal level is not a surprise.

(Laughter.)

MS. WEISS: What success looks like.

(Laughter.)

MR. BRAUN: Right.

So I think it needs to be understood, and I think the consortium, the successful consortium, has to show at least an understanding of these issues and how they propose to deal with them.

Final thoughts, in the last minute and a half: I do believe that these important

federal dollars should be invested in building capacity that can be leveraged over time. So there is going to be sort of a one-time aspect of it, but it should be leveraged so that its value plays out over the generations.

I agree with the others that we ought to be supporting alternative strategies, and the considerable cost, both from the federal side, but also the state contributions, should be more tied to what I call dual-use. So, if we are thinking about this teacher marking it and moderated marking, that is great. By itself, I don't think that is a wonderful form of teacher professional development. But if we extract from that activity, teachers go back and say, "Let's take three or four of these extended projects" or "Let's look at what really went on. How are they linked to instruction? How can we do a better job of preparing for students for that?", that kind of professional development I think is really valuable and leverages the investment you are already making.

The same way, we have already talked about the release of high-security items into the less-secure, open, diagnostic assessments is another example of dual-use.

I think we want to encourage development of educators at all stages of development. That has been said more times than we need to. I think we need to build in requirements for formative evaluation of the assessment system and with independent audits.

It is really important, I think, that the notice encourage states and consortia to do the right thing. One of the problems, I believe, of No Child Left Behind, that it encourages states to do the wrong thing in some cases. And at the same time, allow states some flexibility on the timing of the adoption, particularly in terms of the relationship to the technology platform.

Thank you very much.

MS. WEISS: Thank you. That was great.

Questions?

MS. JONES: So could you give us a little example of what formative evaluation of the assessment system would look like?

MR. BRAUN: Yes. I think that, when you think about an assessment system, the first issue is, it really goes back to the point that Scott made, that is, what is the theory of action? And how does the assessment system you are proposing align with that theory of action? How is it going to help?

And I would argue, and I think others around the table have argued, that this notion of articulation and of coherence is essential. So our first thing should be, how do the prerequisites, the second or third slide, the prerequisites for an assessment system, are they in place? Then how does the assessment design that you are proposing, how does that match up with the standards?

I think too often now, when we talk about alignment, what test vendors tend to do is to say, they draw lines: here's a standard. Here's an item that meets the

standard. So, in the end, you see, yes, every standard is matched to one or more items. Every item is matched to one or more standards, and so we are aligned.

But when you look at the academic side, when they talk about breadth, depth, and so on, they typically find substantial gaps, particularly in terms of what my friend Jim calls the HOTS, as opposed to the LOTS.

(Laughter.)

So we want lots of HOTS.

(Laughter.)

So I think looking very seriously at the degree of alignment and the coherence across grades, and to what extent they really represent research-based developmental trajectories, I think would be, for me, probably the most important part.

Then you have all the sort of the regular stuff. You know, does the psychometrics work, and does it work in a reasonable way? And what are the reliability and sort of predictive validity coefficients?

So there is going to both, what I would say, an in-depth qualitative analysis together with empirical work, and that takes time.

MS. WEISS: Yes, go ahead.

MR. NELLHAUS: Thanks.

You described this four-component system, and you contend that it will reduce the value of inappropriate preparation. So I was wondering if you would talk a little bit about, one, what do you think of as inappropriate preparation? Why do you think this four-component system will minimize that?

MR. BRAUN: Okay. Well, inappropriate preparation, let's say in math, since I come with a math background, would be drill and practice on specific item formats without focusing on sort of a higher-level thinking.

So, for example, we have, even today, lots of examples of students who are taking advanced algebra course, so-called advance algebras, come to do the SAT math and

score very poorly. What one finds is that they have no strategic thinking. That is, they can do the problems at the end of the chapter because they know what that chapter was about and they can apply that.

But if you give them a problem which is not clear what strategy to use, and in what order the strategy should be used, they are deficient because they haven't been asked to do that.

So I would say, in terms of our goals for mathematical power, that kind of drill and practice on chapter things would be representative of inappropriate test preparation.

MR. NELLHAUS: So how would this four-component system mitigate that?

MR. BRAUN: Well, you know, I am going to be very deep here. I am just saying that, by having these different items, and not just different items, but by really getting students to think hard about real-world problems, so these extended problems, and even

ones that are the on-demand with preset materials, are meant to stimulate broader thinking and showing, for example, how mathematical modeling is related to real-world problems that you don't get from the typical multiple choice problems that you see on so many of today's exams.

MS. WURTZEL: So I would like to ask a clarifying question that follows up on Jeff's.

MR. BRAUN: Sure.

MS. WURTZEL: Which is you have this four-component system, diagnostic, extended projects, on-demand, and then on-demand standard. So do you see all of them as part of an accountability system? Do all of those roll up into a summative score?

MR. BRAUN: They might.

MS. WURTZEL: They might?

MR. BRAUN: But not necessarily.

MS. WURTZEL: Okay.

MR. BRAUN: And again, it would depend, partly it would I think depend on what

your developmental pathway is. In other words, a state or the consortia may not feel that they are ready to roll them all up into the sort thing that counts immediately, but they might be on a pathway to say, okay, let's see how this plays out, particularly if they are going to use the technology platform because things can go wrong.

So, once they have developed enough experience and confidence that the platform works, then they might be ready to say, yes, let's fold that into the accountability mechanism.

So I think there should be some flexibility in terms of how these play out, but I think where I would minimize flexibility is in the imperative to expose students to real-world situations in which they have to, whether it is doing the math, writing essays, argumentative essays, doing scientific experiments, et cetera, et cetera.

How that plays out over time in the accountability mechanism, I think a little

flexibility, given differences in capacity, differences of technology, experience, would not be unreasonable.

MS. WEISS: Yes, go ahead.

MR. BRAUN: Yes?

MR. WISE: Henry, I thought one of the points you made at the very end about the need to sort of let -- that capacity-building could be leveraged over time was a very important point, but can you say a little bit more about how that could be done?

MR. BRAUN: And can I quote Laurie Wise on that, like I wouldn't like to talk about that now?

(Laughter.)

Well, for example, I think if you have a technology platform that supports, as most would now, supports web-based assessment, then it is also going to support web-based conferencing, web-based online courses, et cetera, et cetera.

So you are not trying to say I am going to spend all this money on the platform

and then use it for assessment. In fact, the assessment use may be a small part of how that platform gets used.

That is why I think it is very important to be clear what we mean by technology platform, what is an appropriate use of the federal funds in terms of building that platform, and how it will get leveraged through dual- or triple-use for different purposes.

MR. WISE: Good. Thank you.

MS. WEISS: So that takes us to my question, which is, what do you think the appropriate components are for a baseline, for the first rung of the technology platform? What do you think are the required elements?

MR. BRAUN: Well, I am not a technology person. I am the only person here who doesn't have a Blackberry probably.

(Laughter.)

But it seems to me that you absolutely need the internet and you need what we call the T1 pipeline. You need the thick

pipes to bring stuff in more or less instantaneously, close to instantaneously, and you need bullet-proof software. It doesn't have to be that fancy. So we don't need maybe the latest version of Windows, which I have heard is not that easy to use, but you need something that is really going to be indestructible.

I know the experience of folks at D.C. that work with technology, they have incredible stories of what happens to technology when it gets into the schools. I am sure, Jeff, you can speak to that.

So I say the bullet-proof aspect is absolutely critical. Less critical, then, maybe some of the fancy stuff.

Now the biggest issue in terms of cost, of course, is who gets the computer? How is that technology platforms accessed by different people? Is it a one-on-one laptop connected to a wireless? Does every teacher have a computer? Are there still computer labs? I am just not expert enough to know

that. I don't believe we need one-on-one computing to be effective for instruction, but you probably need a pretty small pupil-to-computer ratio.

MS. WEISS: But in your picture, what is the role, what are the different roles technology is playing? You were talking about maybe it is for creating assessments, for delivering assessments. It is video conferencing. What are the different functions that you think it should serve?

MR. BRAUN: Yes. Well, it can do any of all of those.

MS. WEISS: Right.

MR. BRAUN: In other words, if you go to the CREST vendors today, they are using technology to create items, to clone items, to bank items, to construct tests, to distribute electronic student materials to a scoring network that is all over the country, if not now all over the world.

They are using the technology through expert systems to score complex

responses, to create all the psychometrics more or less automatically. So there is an incredible use.

Is that something that the states should be doing? Or is it more sensible to say those states should be partnering with those vendors, for-profit or not-for-profit, and leverage their investment in an appropriate way?

That is why I am being necessarily vague around the technology platform. I think it depends on the nature of the partnership among the states and, then, among the states with whoever their vendor or vendors will be. That should govern, to some extent, what investment the consortium is proposing in terms of their own technology development.

MS. WEISS: Any last question?

(No response.)

Great. So I think we now get to turn to the discussion part.

Thank you. I did receive questions from a bunch of people in the audience.

Rather than just ask them individually, I think, hopefully, you will hear many of them woven throughout the conversations that we are going to have.

The question that I wanted to start with is the big question that we at the Department have to wrestle with as a prerequisite to all the rest of these pieces, which is something that many of you touched on, but I don't think that we walked away having a real firm picture in our heads of whether there was consensus or not and around which areas.

That has to do with the question of, how many assessments in this assessment system do you think we need and why? What role are they serving?

We heard a lot of pieces of it, but I am not sure that I could walk away and say: here's what people think.

So, yes, go ahead. We will treat this more as a conversation, so feel free to just kind of launch in.

MR. WISE: Okay. Well, it seems to me, if you only wanted one end-of-course summative or end-of-year summative assessment, you wouldn't have to invest all this money. You have already got that.

So the real opportunity to advance things is to figure out how to intermix interim -- and I use the term "interim" sort of loosely to mean not the end of the year -- assessments and collect summative information.

Exactly how many you need I think will depend on a careful analysis of the curriculum and of the objectives to see what the sequencing of instruction might look like, how many --

MS. WEISS: How many times a year the assessment interim thing is given, right.

MR. WISE: -- times a year you would want to administer it. I guess I am not as shy about trying out mixing, you know, using some of the interim information as part of the summative measure. I know others have said, oh, keep the formative and the summative

separate or people will start gaming it.

But if you have enough different pieces, people will stop trying to game everything. The important thing is that it provides useful information, and useful information that people can act on in a timely manner during the year, rather than -- I think Gary was very graphic in pointing out the problem that we have now, which is you don't get the results in any timely manner.

MS. WEISS: Go ahead.

MR. MARION: Well, like Laurie, I sort of evade the direct question, or just 3.5.

(Laughter.)

I mean, what is the right number?

MS. WEISS: Well, thank you.

(Laughter.)

MR. MARION: And that is sort of an average over grades and subjects is the problem.

(Laughter.)

I just have to figure out the right

component.

So I don't think there's a single answer, but, like Laurie, we already have the once-a-year. So we have got that nailed.

MS. WEISS: And that is still in the vision?

MR. MARION: That is still in. I think everybody said it. Gary's was a little bit different than that, but I think you could argue his spring one is close to the end of the year. It is around there. So at least we all have one.

I would say, just to be clear, I think that formative should be daily, weekly, and that's, no offense, I don't want the Department, I am not even sure I would want the states, dabbling too closely. I want support for that. I want professional development support, but I don't even want you looking at it.

But the interim I think needs to be frequent enough that it would garnish some attention and serves as a vehicle for sort of

testing important, big ideas, but not waiting until the end of the year.

So I would argue that a minimum of twice a year, once a semester beyond, so that would be three assessments. I would say at least twice. I can make the case that, as you either get up in grades or in different subjects, you could go as many as four or six of them, but then you start running into cost and other kinds of constraints. But more than just one other time a year.

MS. WEISS: And are these all performance tasks to you guys?

MR. MARION: It depends who you ask.

MS. WEISS: Or are these interim ones? Because it sounds like some of you think that they should be measuring whatever was recently taught.

MR. MARION: Much more on performance task approach that is really focused on this. But I did hear some things where I could be -- I like Laurie's idea.

Every time I hear from Laurie, I like it. You know, that you could actually mix it with a piece of an adaptive assessment, where you could get some broader information, along with a performance task in there.

I keep talking over Jeff.

MR. NELLHAUS: Getting back to how many times a year these benchmark or interim assessments would be, I would argue, one, I will still contend they should not be part of the accountability system. I would also make them optional. I would not require them because in too many of our schools there's already too much testing. We are getting a lot of complaints of the testing pushing out instructional time.

Where I would require them is in our accountability system, as we are identifying schools that are underperforming, then we can require them to do this. Because the idea of the benchmark assessments is to improve curriculum instruction in those schools who need to improve curriculum

instruction. So we can gain economies by not requiring them of everyone, and we can offer a tool to schools that need the tool.

MR. DUECK: When I take a look at testing, I generally see three purposes. One is the accountability function, which has had a lot of discussion. We need to make sure that the taxpayer is feeling that, indeed, they are getting what was being paid for.

I know there is the improvement function. There has been a lot of discussion about that.

Then there is a third element that I think sometimes gets lost, and that is the fairness-to-students element. You need to have the guarantee that there is consistent interpretation and application of the standards, regardless of gender, location, background, or whatever it is.

So when I take a look at trying to find a test that is perfect for meeting all three of those, I don't know that it really can happen. So, therefore, it depends where

you are going to put the emphasis on the right syllable.

(Laughter.)

What I would suggest is that, from the national perspective, accountability is what really is reigning over the improvement function, and fairness to students has to always be applicable.

So, therefore, for me, it is an annual assessment and then having the various LEAs or schools take a look at the issue of the formative assessment on a far more frequent basis.

MR. COOK: I would say ditto, but what Henry's said and Laurie said I think is really good. I want to get back to what Scott said, and Ron Hamilton has been thinking about this for many years.

What information do teachers and principals need to help them help their kids prepare for college readiness, and how frequent should that be given to them?

I mean I think the idea that we

need to design a system so that we have the end-users in mind and guide them in the direction that we want them to go is the answer to the question, "How many?" I mean I put three. Some put -- I don't know.

But what's the point? Isn't the point to give parents, and I think students because I think, in my view of formative assessment, student self-assessment as well as interim assessment, student self-assessment needs to be hard. Students need to begin taking responsibility -- or, no, we should begin giving students the responsibility for their learning and provide them with tools and mechanisms to help them recognize where they are at and where they need to go.

But I think, in answering the question, "How many?", I think we need to ask the question, what's the frequency and what kind of information do we want to provide to parents, teachers, and administrators to use to help their kids? That is the question.

I think the notion that we need --

the end-of-year assessment, the one-time, 65-item, 70-point, constructed, multiple choice thing that we created as an artifact of ESEA or ISA and then NCLP, that may not be helpful for students, parents, and teachers in helping their kids.

We need to think through this idea: what do we want to provide to the educators and parents and students? And how frequently would that be useful for them to make judgments about their students' progress toward college readiness or career readiness?

MS. WEISS: So what is the argument for having that end-of-year test that everybody says, "Yes, that is part of it because we do that now."? We've got that one. Let's talk about the other pieces. Let's talk about that one. Let's talk about that one and just make sure we understand what the purpose of that one is and should be, and what it looks like a little more.

MR. COOK: To keep them from going on field trips at the end of the year.

(Laughter.)

MR. BRAUN: I think it represents, for many people, I think it will represent an anchor to the system. In other words, it represents something that they feel comfortable with and, to the extent that it is used in conjunction with these other perhaps more innovative, some would say wrong-headed, innovations, it provides a certain measure of confidence. I mean that is sort of, if you will, political/psychological role. But in terms of pure measurement --

MS. WEISS: Yes, that is what I was looking for. It has got a psychometric role, yes.

MR. BRAUN: Yes, it is going to represent, because of the way we do it now, it builds in a lot of reliability. But it should be reliability, not at the expense of validity. We could do a lot better job.

Jim's examples -- and there are many in this country, I am glad to say -- are examples of really well-done, objectively-

scored assessments. As we move to a computer-based platform, that is going to increase enormously. So, with that, we will be able to get both validity and reliability.

One of the things that concerns me, I think I feel we are a little bit out of balance. One of the problems with summative assessment, particularly sort of externally-mandated assessment, is that, as Bob Mislattie likes to say, they are sort of dropping from the sky. You know, they drop in. They scare everybody. They come out with the numbers, and the entire interpretation is based on those numbers. They are completely decontextualized.

When we come to formative or diagnostic assessments, intended to help instruction and learning, let's not pretend or let's not believe that that's all the teacher has to go on. I mean it can be a very important source of evidence, but good teachers are watching the kids all the time. So they have a lot more context.

What we really want to do is we want to make sure that they are getting the best kind of information based on research, that in conjunction with classroom observations and their knowledge about the student and what the student's current homelife is like or this student is getting bullied, so they are not paying attention, or whatever. But they are making certain judgments.

I don't think that we, even at the district level, can tell teachers everything they need to know. We have got to rely on their professional judgment.

What we have to ask is, what is it that they most need that would complement what they learn day to day in their classroom that would give them the most comprehensive picture of what they need to go to the different instruction.

So I think we need to look at it in the context of teacher, not in the context purely of one kind of assessment versus the

other.

MS. WEISS: So let me ask about student growth. This is something that was in our notice and it got sort of touched on a little bit here and there. But, within these, how do you think we measure and start looking at the issue of student growth?

MR. WISE: Well, I think looking at student growth is sort of absolutely essential, both to communicate better with parents. Here's where the student was at the beginning of the year. Nine months later, is the student better off, and in what ways and how?

A problem has been if you have just the end of the third grade and into the fourth grade, and there's not much connection between the third grade curriculum and the fourth curriculum, then you are comparing two numbers that aren't all that comparable.

So that is one of the reasons that I am quite hopeful that, with a better articulated set of content standards across

the grades, we can talk about growth or the value that a school adds to a student's development in terms of a sequence that makes better sense now than the current sequence of end-of-course tests, end-of-year tests.

MR. COOK: I work on growth. That is one of the things I look at. I agree with Laurie. But there are technical issues and challenges to doing that.

I think, if we have an articulated, well-understood trajectory and pathway or a series of pathways, then we can begin to make statements about progress and the relationship between a student's progress and some sort of institutional or individual effect on that progress, but we don't right now.

The concern I have, and have had, and I like value-added stuff; that is what I do. But the problem I have, and the concern I have, is when we provide metrics that have units attached to them like classes or teachers or schools or districts, then what happens is the users make inferences about the

relationships between those numbers and the effectiveness of teachers and principals and students. Those inferences may or may not be valid, but, certainly, probably aren't causal.

That is why I like some of the stuff that Scott and Damien and other folks have done at the Center about looking at sort of this notion of normative growth. When we can begin thinking about the sort of patterns of where children are or schools are, and then start modeling that, and then making statements about where it seems you are at relative to some population, I think that is a lot better.

The concern I have right now is -- and I think growth is critical; I am not going back on that, but we need to know more about this trajectory and what it looks like. I think that first needs to be established before we start making statements about you are making adequate or not adequate growth or this school is growing kids at a good rate or not. Because, quite frankly, I don't know if

we can make that kind of causal inference from what we know.

MR. MARION: I think actually just the four of us spend a lot of time thinking about this. It was interesting that none of us actually mentioned it that much. Because, first of all, we are starting to stay away from accountability, and that really is more the accountability mechanism.

But more to the point I think is -- I will speak for me. Henry and I have been working on a paper together, so I get a little bit of this.

But we have been doing growth, and people have developed some pretty sophisticated, incredibly sophisticated, analytical models to try to get at what Gary is talking about, trying to make causal inferences where really none exist. They have been doing it on some not-very-good assessments.

So we have the analytical tools to measure growth. What a lot of us are saying

is let's get good assessments, based on good whether it be learning trajectories, or whatever you want to call it. We could figure out how to do the evaluative modeling, but these foundations need to be there for us to be able to draw proper inferences to say that a kid grew four points, so what? What are they growing on that a lot of us care about, and the measurement aspects are a more important focus.

Now, if you have got a bunch of economists and statisticians in the room, they might feel differently. You happened to pick some psychometricians, so we tend to care more about the measurement piece of it.

MS. WHALEN: So can I ask a follow-up question to that? Besides the learning trajectory, what are the preconditions in an assessment that you would think that should be required to enable systems to do that?

MR. MARION: Henry wanted to talk anyway. I know he can answer this question.

(Laughter.)

MR. BRAUN: I'll do it.

Let me sidestep that for a minute.

(Laughter.)

MS. WHALEN: You guys are very good at that.

MR. MARION: We could answer this one.

MR. BRAUN: I think it is important to distinguish the individual from the educational unit.

MS. WHALEN: Uh-hum, right.

MR. BRAUN: So our first goal is to be able to generate information, evidence, about a student's progress. Then we need -- and that partly goes to your question, Ann --

MS. WHALEN: Uh-hum.

MR. BRAUN: Then you need a set of content standards and a developmental model that says: here are the typical ways in which students move through this domain.

Then your assessment needs to have the measurement strength, if you will, so that you can, in fact, make reasonably reliable

judgments about what path a student is on and how far along that path they are, so that you can say something about their growth.

I agree that the numerical scales we have now, they tell us a student has gained by eight points, but you can't say what that score meant then, and you can't say what the eight points more mean now except that it is eight points more. Who knows what that is? If we are talking about vertically-linked scales, we really don't know what they mean.

So I think this, to me, is one of the really fundamental goals of the innovation, is to really build assessments that support inferences about individual progress along domains. And I laid out some of the preconditions.

I think that needs to be made very clearly distinct from what goes on or might go on in accountability, where we take these individual --

MS. WHALEN: Right.

MR. BRAUN: -- performances,

whether status or growth, aggregate them to either the classroom, the school, or the district level, and then want to make inferences about the effectiveness of the classroom teacher --

MS. WHALEN: Right.

MR. BRAUN: -- the school or the district with regard to student learning.

We have sort of at least alluded to the enormous difficulties, the technical difficulties, in trying to make those kinds of inferences, which are really causal inferences, from non-randomized tracks. You know, there's all sorts of selection.

The paper that Scott is referring to, we are on a committee, an NRC/NAEA, which is trying to look at what people are doing around value-added, what are the concerns, and how might an accountability system employ value-added.

The distinction between growth and value-added is that, in a value-added model, you are explicitly trying to adjust for

differences among students across units, so that they are more or less on a level playing field. That is a good thing to do. It is hard to do.

But I think we need to separate the basic, I think more fundamental, issue for this notice, which is to build good assessments to support student learning and to tell us where students are. If we do a good job there, we will be in a better position to do the kind of accountability that is surely coming down the road.

But, as I said in my little presentation, it is not a guarantee because there are lots of other problems.

But I worry, if we commingle them, that we are going to really get more confused than we are.

Did I answer your question? I can't even remember now what it was.

(Laughter.)

MS. WHALEN: It as about the precondition

MR. BRAUN: Yes, well, precondition.

MS. WHALEN: You named a handful of them. I guess, to your point, how can we be explicit, more explicit, about what we would want to ask in an assessment to allow for the type of accountability that you were just referring to?

MS. WEISS: At the student level.

MS. WHALEN: At the student level, right.

MR. BRAUN: Yes. Well, I think the preconditions I have would be sort of a core set, but I am sure people around the table would have examples of other things they would think as sort of, if not preconditions, as sort of -- if not required, it is at least good to have. There might be other combinations that would work just as well.

But I think the onus would be on the presenters, the applicants, to say why this particular set of conditions that they either have in hand or propose to do would be

sufficient to support the kinds of inference around student progress that is being asked for.

MR. MARION: But I think to specify in the notice, what you could ask for -- and Henry is not being vague about this -- is that you can say, how does the system you propose allow you to make inferences about students' progress in a domain moving from, say, fragile understanding to deeper understanding, or something like that, but specify that it has to be sort of domain-specific progress that you could then judge for an individual student.

I think if people could propose a way to do that, then that, in fact, gets at they would have to satisfy a lot of the preconditions that Henry is talking about, and issues of growth and scale that we are talking about, because that is really the core of what you are asking about.

MR. NELLHAUS: I just think the one thing that comes to mind when you mention

measuring growth to people is that they automatically believe your assessment system needs a vertical scale. So I would hope that the notice that you put out, since we know we don't need vertical scales to measure growth, that it is not a requirement of the RFP.

MR. MARION: But it shouldn't be ruled out entirely.

MR. NELLHAUS: No, no. It is just not a requirement.

MR. COOK: I just want to highlight what Henry and Scott just said. I think the issue is the inference. We want to make sure that we support the inference of a student's progress in a domain. That should be the target.

I mean, if we get that right, then trying to sort out how we monitor growth over time and scaling that to make meaningful statements about progress kind of follows. We do know a lot of things about how to do that, but let's get the inference right first.

MS. WEISS: Okay. I just want to

ask one more time, so if you are the ones answering the RFP and saying, "Here's how what I propose today gets the inference right," what are the key things that would allow you to argue that what you are proposing is a system that would allow us to make those inferences with some level of validity?

In other words, it is one thing to sort of push it off on the people sitting in this room who have to respond. It is another thing to give them more advice, so that when they are responding, they have got the ideas and the concepts of what it is that --

MR. MARION: Get your pencils out. Laurie is going to give you the answer.

(Laughter.)

MR. WISE: No. I am first going to talk, just briefly, to the Department people.

I think the two most important things that you need to evaluate in general terms is sort of, what is the potential value of what is being proposed, if it works? Then, what is the likelihood that it will work?

I am not convinced they have to be on an interval, ratio scale that allows that kind of transformation.

So, clearly, bringing together evidence that this has been done in the past and citing evidence is one way of both enhancing in the proposal the perception that it is likely that it actually could work because people have done pieces of it, and it is also another way of, it can also be used in, I think, portraying what the potential value or payoff from this improvement to the assessment system might be.

Now that is without getting specific at all, but I think you need to be not too specific in exactly how you get there, but you need to figure out how you can fairly and evenly evaluate the value and the feasibility of the different things that people will propose.

MS. WEISS: Great. So that is the advice to us. So now, if you are giving advice to these guys, what would you say?

They are nodding and saying --

MR. MARION: Well, cite the evidence.

(Laughter.)

I am not afraid to put a stake in the ground, as you see.

So I think, first of all, the states are limited because -- well, I just want to clarify, too. I just got a phone call yesterday, and somebody was surprised. I said, no, I think the notice says the consortia have to adopt "a common set of standards", not "the common standards". Correct?

MS. WEISS: That is the article we used, correct.

(Laughter.)

MR. MARION: Okay. All right. Now who is being vague?

(Laughter.)

MS. WEISS: I mean that's right. That is what it says, "a common set" --

MR. MARION: That's right, "a

common set of standards".

So NEECAP states, you could just rest easy.

But the thing that I would argue is so, first of all, the states are somewhat limited by, let's just say, the common standards, and if they don't do a good job of articulating the content. But let's give the benefit of the doubt and say that they actually do a good job of building cross-grade coherence.

Then I would argue that, as part of a proposal, I would want to see the state -- I am guessing that they won't be at a fine-enough grain level to do all the stuff we want to do. So the first thing I would want my curriculum and content folks and learning folks to be doing is sort of figuring out perhaps how to specify this out a little more. So give me information about what Laurie was talking about, these within-grade trajectories.

Then, within that, I would -- of

course, I gave you my rules -- I would build these units around key points in that place where, then, I could say, you know, if the kids are satisfying and demonstrating competence, mastery, whatever you want to call it, at this unit that I am doing at the end of October, and they have given me sufficient evidence that they understand this, that tells me that they are somewhere along that trajectory.

Now if I am smart about how I build my end-of-year test, and we brought in the token foreigner, but he is still from North America -- (laughter) -- but we can reach across the oceans and look at Australia, New Zealand, the Netherlands, where they actually are building some really nice assessment systems, based on this notion of learning progressions, to give us evidence about how people are moving along a growth continuum and learning progressions. Particularly Australians have a nice way of blending assessment and curriculum that way.

So, then, I would want to say this is now a content-based, a construct-based way of thinking about growth.

MS. WEISS: Okay.

MR. MARION: So they can have a little more time to write this than my five minutes to think about this here or five seconds. But that is the way I would go.

MR. COOK: Just to follow up what Scott said, you have to put a stake in the ground about scope and sequence and commonality. Then the idea, I like the idea that two to six units, within each of those periods that you choose to have the assessment.

I mean a lot of that is the challenge. You have this end-of-year examination, whether it is the beginning or the end of the year, and everyone has got 57 different ways of trying to get there. Part of the problem is teachers don't know which one is best, better, or even productive. Then they choose textbooks and materials to try to

get there, if they are doing that. We need to have that sequence provided by the curriculum people to get to the point.

MR. WISE: I think one of the things that is important for the states to do, as they are putting something together, is to be as explicit and clear and detailed as possible about the theory of action or theories of action, of what they are hoping to accomplish. Then break it apart step by step and say, "And here's the reason you should believe that this step will actually work."

So, for example, I was really interested in Henry's point about leveraging because I am, and I know a lot of states are concerned they will get this money one time. They will develop these really whiz-bang assessments, and then they won't be able to afford them.

So can you build a rationale that says, by investing in computer-based platforms, actually, once it is up and running, you eliminate printing and shipping

and scanning and a lot of costs. So that you have actually reduced your ongoing costs with these upfront costs. You have leveraged something really concrete.

So make a good argument about that and then why it will be sustainable. Because the feasibility is at least as important as the value-added.

MS. WEISS: Thanks. Okay, so let me change gears for a minute.

We touched on this a little bit when we talked a little bit, I think, when Gary was talking about some early childhood stuff, but I want to just ask specifically the question about, if third grade is the entry year at which we are doing assessments, are there any specific things that we should keep in mind about those third grade tests that may be different from other tests that we are creating throughout the spectrum that will help those tests signal to kindergarten, first, and second grade teachers what it is that they need to be doing to prepare this

kids for success in third grade?

MR. NELLHAUS: I would just suggest that, in terms of first grade/second grade, we could develop a series of benchmark assessments, low-stakes assessments, that would build toward the expectations in third grade.

So I understand the issues of development. Students are developing at different rates at the early grades. So we certainly can at the individual level look if they are making progress. So I think that would be the best way.

Certainly, using the results of the summative tests that third grade schools are looking at, what are the areas that they are not teaching effectively, if their students are not performing well on the third grade test.

So, again, I see the system working. Maybe not summative tests at the very early, but certainly a series of benchmark assessments would help a lot of

schools.

MR. COOK: I am going to follow up on what Jeff said. I don't think it follows, just because you start assessment, summative assessment, in third grade, that formative and interim assessments previous are of no -- I mean you have to have that in place. I think you have to have the formative and interim assessments, and, yes, you are going to have different trajectories. So, if a child in first grade isn't doing more than two-letter words, you are not running around pulling your hair out, but now you know where they are at.

So I think the notion that just because we are concerned about assessing before third grade, and the inference of what the score means, you still have lots of tools and guidance with formatives and interims prior.

MR. MARION: Just one quick thing. I actually think you should ask this in Denver of Laurie Shepherd and then duck.

(Laughter.)

I would argue you need to stay, I think really stay out of the early grades. I mean to make a totally perhaps invalid causal leap -- (laughter) -- I mean if you look at where we stand internationally, right, we are doing really well on international comparisons in fourth grade; we get progressively worse.

I mean I think there's a lot of good stuff going on in K-3, and I don't know that we want to screw it up.

(Laughter.)

A lot of those, they have good early reading assessments that are really focused. I mean we have some concrete markers. The kind of stuff that Laurie is talking about, about getting a good sense of curriculum and content, we have that in the early grades. We know we have to leave kids a little bit about --

(Laughter.)

MS. WEISS: We are not even going to ask what the causal leaps would be for what we, therefore, should do in grades 4 through

12 that would put all of you out of jobs.

(Laughter.)

MR. WISE: In California, there's a big effort to look at and try to do something about the achievement gap. Most of the evidence says it is already there by the third grade.

Well, the idea that the K-3 education is functioning and is doing what we want it to, I think is challengeable with regard to equity for students of more impoverished backgrounds or lower socioeconomic status, and so on.

I also work some in organizational theory. One of the things about organizational theory when things work is when, as a manager, you put in place metrics that allow you to track how well you are doing what you are doing, and think of ways to improve both the metrics, but also what you are doing.

So I am a first grade teacher, and if there is no metric that allows me to see

how effective am I being, and where I could be more effective, I think we have lost an opportunity for continued improvement.

I mean there are metrics there now. It is the principal with some more subjective feedback, but --

MR. MARION: But don't you think the principal is going to say, if the kids aren't reading in third grade, that we can't wait until third grade to do something, that they have actually, hopefully, put something in place?

MR. WISE: It would be nice if you could tell for sure whether it was the first grade or the second grade teacher that messed that up.

MR. NELLHAUS: I would just say that, we have had how many years of Reading First? And certainly all the states were required to do benchmark formative assessment as part of the Reading First program. So I would ask the USDE to look back and see how successful that requirement was in actually

improving the programming.

We had an interesting situation, I think, here in Massachusetts where we saw students making progress on Dibbles and on GRADE, but not doing as well on our own state assessment. If you looked, they were not in alignment. Dibbles and GRADE put an emphasis on decoding, and our State test puts an emphasis on comprehensive. So teachers in first and second grade were focusing on decoding skills and not on comprehension skills.

So it talks, again, about this need for a unified system. If we are going to have benchmark informative assessments, they need to be anchored in the summative assessment and what is going to be valued in terms of the accountability system.

MR. COOK: And let's not forget the point of common standards is the common standards are going to go down to kindergarten, and the sequencing associated with that is going to go down to kindergarten.

That is where it needs to happen.

MR. MARION: And notice that I didn't say "Dibbles" when I was talking about --

MR. WISE: And from a learning trajectory point of view, both decoding and comprehension are important. People would argue that decoding has to come first, so that you have something to --

MS. WEISS: Okay. Let me turn and ask you about the role of the district. This is something that we talked about, I think, at the very opening, but didn't really touch on a lot other than just sort of referring to teacher involvement in both development and scoring of the assessments.

But what do you see as productive, good roles for LEAs to play with the half of the funding that flows down to them as part of a grant like this? And what do you think would be roles they shouldn't be asked to engage in?

MR. COOK: What is the connection

between an LEA and an SEA who is in a consortium? I guess, is the presumption -- and maybe it is my misunderstanding -- if a group of SEAs form a consortium, does it follow, then, that the LEAs and the resources that are going out for the other half of the money is going to LEAs? Or is it just general to all LEAs?

MS. WEISS: No, it is to LEAs in the states that participate in the consortium.

MR. COOK: In the consortium, okay.

MS. WEISS: And it is not all LEAs. It is the LEAs that choose to participate. So the question is, what is it that LEAs would have to do so that LEAs understand whether or not they would be interested in participating?

Remember, if the test becomes the state test, everybody will take it. So this is the development, this is the question about the development of the tests and the role that LEAs could or should play in anything from conception up through field testing, I guess.

MR. BRAUN: I think all of us spoke

eloquently --

(Laughter.)

MR. MARION: Speak for yourself.

MR. BRAUN: -- on the role of teachers in development, in assessment development, grading, and so on. And I think that the LEAs play an important role in sort of organizing that activity.

Again, in terms of the dual-use, figuring out how to leverage that into extended professional development that goes beyond whatever is being required for the assessment.

So I think the LEAs could really pick up on the professional development side. I was saying at lunch many districts are developing what they call professional learning communities in these collaborative efforts across a school that include teachers, specialists, and so on.

So thinking about how you link the activities of the teachers who are participating in marking or auditing other

people's marking, how they can bring that back into a professional learning community, that is something that seems to be tailor-made for the LEAs to be organizing on a district-wide basis. So that is just one example of what they could be doing.

MS. WEISS: I haven't heard a lot about teachers developing assessments. We talked a lot about marking and auditing and the PD side, but what are your thoughts on the development side?

Go ahead, Jeff.

MR. NELLHAUS: Well, actually, in terms of developing performance tasks, we have conceived of this idea in Massachusetts, that we would actually make a competition and have teachers develop them and submit them to us, have them peer-reviewed, and have them eventually tried out by other schools and eventually folding them into wider use. So I think teachers could play a very strong role in that sense.

I would also say, in terms of just

designing it, we talk generally about performance tasks. So a task to enhance the validity of the assessment program, of the on-demand test, that is really the reason for doing that.

In just thinking about how to do that, I think states would do well by bringing together forums of teachers and school leaders to think about just how to do that, all the logistics, the length of the tests, the nature of the tests. I think we could benefit greatly by bringing them into the actual design of the process.

MR. MARION: So there are certain things that you had in The Federal Register that I wouldn't do, like paying for pilot testing and things like that. But there are certain things, so I will agree with Jeff that you can work with teachers to help develop performance assessments, but my experience, and pretty extensive experience -- we did this in Wyoming. The Wyoming body of evidence, activities consortium has developed a couple

of hundred really complex performance tasks that have been pilot tested and validated, but teachers couldn't do it without really good expert leadership and guidance.

It is really hard to develop, especially if these are going to have uses beyond just the classroom. So this serves as a good model.

But we had a tremendous number of teachers working with us on development and then, more importantly, piloting these things, and then examining student work. So it is tremendous form of professional development. It was a tremendous of form of assessment-building, assessment development.

But I would put a lot of resources into that. I think, again, teachers in the consortium, if you had a consortium of states, content area of teachers at particular grades could be organized. So have some of this money directed to the LEAs to pay teachers for these sort of extensive summer institutes where they really build curricula units,

again, with a lot of sort of keepers of the content to make sure it does follow the standards, et cetera.

But that is the way that I would use teachers, to do them in real work. I think the idea of having teachers score is good, but I actually am much more interested in them scoring locally and then getting feedback on how well their scoring is going. So they are doing that as a regular thing.

I mean it is sort of like when you teach elementary stats, you know. You teach them the standard deviation by hand, and then once they get a calculator, they throw it at you because you can't believe you actually made them do it.

(Laughter.)

So I think there is a certain Astonian effect of scoring all these things. So you actually gain a lot of skills.

So I would use them on the development end. I would use them on development, in my model, development of these

units, and then piloting some stuff, and then reviewing student work.

MR. WISE: If I could just emphasize one thing that Scott just said, for teachers to really be effective in developing good questions, and so on, requires a certain amount of training and monitoring and feedback. But there would be a tremendous value, I think, in investing in the training of teachers in how to create evidence-rich test questions and exercises, and how to create and follow scoring rubrics in a valid and reliable way. If you have done this investment during the development period, you then have a cost savings down the road in a huge cadre of teachers who are well-trained and can do this. You don't have to pay sort of the going labor rates for item writers, and so on.

I wouldn't have teachers do psychometrics though.

MR. NELLHAUS: The real way to get the enduring benefits of that is to have that

included in our teacher preparation programs.

MR. COOK: Yes. I mean, just back up with what Laurie said, and some of what Jeff and Scott said, I think LEA should explain how they are going -- and I think all of us or at least I thought we heard some consensus about that one summative is probably not a good thing. Probably there needs to be interim and formative included in some fashion.

How is the LEA going to, one, I think support the development of the interims and/or formative structures? I am not going to call it "formative assessment" because I agree it is a formative instructional process. How the LEAs are going to participate in that, how the LEAs are going to set up a professional development program to implement the interim and formative assessments that are being developed at the state level across the consortium, maybe how they are going to use funds to communicate not only within their district, but across districts, and maybe

across districts, but across states on how they are implementing that. And then how they are using the results from the interim and summative tools to formatively evaluate their professional development and formative assessment process.

I think you need to have all of those pieces in for an LEA to be a part of the the grant, because I think Laurie touched on it, and I think in what Henry said. You need to develop a process so that the LEAs are expanding and sustaining capacity and the implementation of the system, and not just this one time you do it real quick, thanks a lot, and we are done.

MS. WEISS: So, Jim, let me ask you, what have you done along these lines? What are teachers participating in? What have you learned about that?

MR. DUECK: As I have been listening, and making sure I understand all the jargon about LEAs --

(Laughter.)

MS. WEISS: School districts.

MR. DUECK: And I realize that. Everybody has their jargon.

One of the things that we are in discussion right now is about having teachers with the low-stakes assessment -- that's grade three, six, nine -- actually undertake the assessment within their own jurisdiction. Within their own school, perhaps there would be a jurisdiction's responsibility to determine that.

We are just in the discussion right now about that as an alternative to the more expensive option, which is bringing them all to the centralized marking center. So I can't indicate to you any outcome of that except this: when we do assessments, and we have a huge data file on all of this, we know that bias is something that works its way in.

Teachers who mark students' paper who they know have different sets of marks, markedly so, than when someone marks the paper that is not known. So, therefore, we have to

be very cautious of that.

We also know that there is bias that comes in because of culture. In a particular community, people have expectations regarding the abilities of students in that community and, therefore, a bias sets its way into that.

That is why what we have always maintained is bringing in marking that is absolutely blind. No one knows anything about the student other than they are a student within the borders of our particular Province.

MR. MARION: Could I just add one thing to it?

MS. WEISS: Uh-hum.

MR. MARION: I think this is a really important question about the role of LEAs. Knowing the slate of experts that you have in the various cities, most of us don't work directly with districts. There are folks that I see in the audience that I know who are in state departments who work both on the support side and working directly with

districts. And then there's folks in the research community, people like out of Ctree like a John Sabovich or people like that who are working out figuring out -- it is the multiple levels of the district. It is from the central office down to the classroom.

So I think it would be worthwhile to get some advice from people who really have expertise in how best to do this.

MS. WEISS: And the states will know a lot of this stuff.

MR. MARION: Yes.

MS. WEISS: So some of it is just giving ideas for them to mull over as they are thinking it through. But yes.

MR. COOK: Why not include the districts in determining how to be a part of -- well, there's the assessment system itself, how that gets incorporated, and then how they want to implement and utilize the system. I think that is a part of the process.

I would presume that, if you are going to implement a statewide assessment

system, you have got all your key districts as a part of the team that are helping you do that.

MS. WEISS: Great. So let's turn to a different question which is maybe more specific advice to us and the Department, about a question that we have been wrestling with a bunch, which is something that a number of you touched on a little bit in your presentations. But it is how specific we need to be or we ought to be in this application; where should we be tight; where should we be loose, and why?

That might cover just sort of a whole host of different issues. So feel free to be specific in this area be tight and this area be loose.

Yes, Laurie?

MR. WISE: So I think you need to have thought carefully amongst yourselves and be real specific on what are the goals of the assessment systems that you are going to fund. Be as concrete, be as numerical almost.

Then, second, you need to be specific on how you are going to evaluate the proposals that come in. So what are the criteria against which you are going to evaluate them? I would put something like some number, limited number of dimensions of value, limited number of dimensions of feasibility as the general idea, and then you fill in the details from that.

I think you want to not be too specific about how to get to the goals because there's a lot of creative ideas out there that we won't, any of us, be able to anticipate all of them in advance. You don't want to shut off the creativity.

But as long as you have got some overriding goals that have to be met and some framework for evaluating the ideas that get thrown at you, I think you will be in good shape, and you will probably end up with a potentially much richer set of proposals than if you start specifying how many items of what type have to go in every assessment. I

wouldn't go there at all.

MR. MARION: Yes, I mean, like Gary said, ditto. I am thinking about the proposal. So that is like the content side. That is the conceptual side.

But I think there is this whole -- we have talked about a little bit the sustainability side and the organizational side. So how is the consortium organized to ensure that these terrific ideas have a snowball's chance in you know what of getting met and getting carried out, and getting sustained?

So I think there is this whole sort of policy governance side that needs to be a part of this. Because we have very little experience in kind of these large consortia that we are talking about here. So I would see that people have thought that through really carefully.

MS. WEISS: Are there specific elements or traits or things you think we need to ask about or look for?

MR. MARION: I would ask about things like, what's the capacity of the varying states participating and what are they willing to give up and who is going to control what decisions? Are they having a central organizing authority? What's the role of the vendor or vendors? Is there an executive director of a consortium? Does it look like a RITA-type consortium and LEA? Does it look like an Achieve Algebra II thing? Or does it look as tight as like a NEECAP kind of thing?

So we have a few existing models now. What models the group of states picks and why, and then have they thought through the various issues where these things could go wrong?

I think you could learn a lot. There's a few NEECAP states represented here. You could learn a lot about how easily this thing could go wrong from the RITA folks.

(Laughter.)

So you get the best ideas, but then you can't get folks to agree on the

accommodations manual, or who knows what, or administration guidelines, and release rules, and things like that.

So there's lots of ways, there's more ways, a lot more ways for it to go wrong than right. So I would want to see that part. It is almost as important as the conceptual ideas.

MR. WISE: Yes, I would just have included this within the framework of the feasibility, one of the important dimensions.

MS. WEISS: Right.

MR. WISE: And a key thing is both the management structure, but sort of what decisionmaking and approval has to happen for the assessment to be implemented, if you are successful in developing it, and what assurances do you have that those bodies within each of the states that are going to have approval authority are inclined to go along with it?

MR. COOK: I mean just take what Laurie said, what he just said. Okay? And

let's make the presumption -- I think it's true -- five years before this thing gets on deck, three to five, I'll say three to five before it gets on deck.

So let's say we get an approval, you say September of next year, is that right? Did I understand you correctly?

MS. WEISS: Uh-hum.

MR. COOK: Okay. So November.

(Laughter.)

MS. WEISS: There's a honeymoon --

MR. COOK: Yes, you're right.

So then we are going to start doing something in 2011. How is the state superintendent in State X, Y, and Z elected, appointed, you know, and administered? What kinds of infrastructure and leadership staff do they have?

So it seems to me you would have to have some governance structure and some commitments through the governance structure that will sustain through administrations. Then you deal with, when you change

administrations, you change procurement structures and procurement priorities.

So how are you going to sustain? What kinds of agreements do you have as a consortia of states to assure that that is sustained?

I think, what kinds of interstate agreements do you have? And how have you talked about interstate agreements? What kinds of intellectual property responsibilities, rights? And all of that stuff needs to -- I agree with Scott, I think the notion of that, sustaining consortium across time is as, if not more, important than the model that you choose to adopt.

A good example in the EOP world is, and I am not trying to be nasty, is the Mountain West Consortium. I mean that was a good --

MR. MARION: A good example of what?

MR. COOK: Of how not to work.

(Laughter.)

MS. WEISS: A good bad example.

MR. COOK: Yes, I mean it is a good bad example.

You know, they had great ideas, but there wasn't as much a cohesive structure. So when the project was over, they had a whole bunch of items that they didn't know what to do with. They didn't have an assessment system.

This could happen the same way, unless you really articulate the governance structure and cross-administration systems in place. You know, you can create something that it starts going really well and it just blows up. I think that is an important piece of the proposal.

MS. WEISS: Can you imagine differentiation across the consortium of different types of responsibilities, especially if you have a large consortium? Most of our examples are pretty small consortia. What if you have a big one? What are the different types of roles --

MR. COOK: I think 20 is big.

MS. WEISS: Yes, I do, too.

MR. WISE: I would think there needs to be some leadership and management.

MR. COOK: Yes, right.

MR. WISE: And it may be that you have got 20 states signing on, but if there is not some either core state or small group of states, or if there's not some external sort of executive director model, I think you are in big trouble because doing everything by consensus of 20 or more folks is not going to move very fast.

MR. COOK: Yes, I mean executive committee, executive board, executive director, you have to have some governance structure in place.

MR. NELLHAUS: It also seems to me that we have two fairly well-established consortia now. We have two fairly well-established consortia now, one on the Algebra II consortia and the NEECAP. In each case, you have a third party actually bringing the

states together with Achieve and with the Center for Assessment.

So the question is, is that something that we have to consider, that there may need to be an entity outside of the various states that helps bring all the states together, sort of a third party without all of the vested interests, a convener? That might be something to consider here.

MR. MARION: But to be fair, the Algebra II, sort of Achieve was really central in convening it, but I think Pearson owns it, is really the vendor now. States who want to get in negotiate with Pearson as well as they have to come to some meetings and stuff like that. But it is a really very different model than NEECAP.

Maybe that is an issue, when you get to a certain size, it does become more of, do you want to buy into this assessment system? This is the one we have.

Because you have to think about that, right? So, if you say there is an

application from 10 states for a consortium, and that is the one that wins, and maybe it is the only one that we are wins. Then what about the 11th state who a couple of years later wants to get in? Do they have to pay back dues or development costs, or how do they do it?

MS. WEISS: Or other states that may have felt like they don't want to participate in a consortium to develop it, but they have go the same standards and want to use something that someone else developed.

MR. MARION: Right.

MS. WEISS: That should be clearly possible in this.

MR. MARION: So it is hard issues. I am glad to be advising and not deciding.

(Laughter.)

MR. COOK: A clear articulated governance structure, I mean that is the issue.

MS. WEISS: Yes.

MR. MARION: Yes. For both

development and sustainability.

MR. BRAUN: I mean I absolutely agree that these issues, organizational issues, are crucial. But just to come back for a moment to content, Ann asked earlier something about sort of like the prerequisites.

So it seems to me that you want the states to be able to lay out their plans for the content development and the standards, all of which need to be sort of articulated and coherent in a way that would support the assessments.

So, even though the money is going to be focused on assessment, so what I am meaning is, are you going to be asking states or are states or consortia going to have an advantage if they are talking about sort of like in-kind contributions? You know, we are going to get this money, but, by the way, we are investing "X" million dollars in building this new set of articulated standards.

So some states might say, well, we

are going to use the money to build the standards. Others will say, well, no, we have a separate funding stream to do that.

So I guess one question is, is that considered to be a bonus for that consortium if they are making that in-kind contribution?

MS. WHALEN: Can I ask, could you conceive of a situation where, within a consortia, states did not have the same curriculum or scoping sequence? I mean they are going to have the same standards, but they may take different approaches to how they, then, implement those standards. So you wouldn't say all of the states have to come together with this common understanding of the curriculum that would be implemented and scoping sequence within a grade or even --

MR. WISE: Yes, I wouldn't say that, although I would think over time, if they don't, it is a great research opportunity.

(Laughter.)

And the states would be crazy not

to continue to try to improve things. If one state seems to be more effective with a particular curricular approach, I would think over time that you would get more similarity in the scoping sequence and the types of instruction that were used to try to bring students along to these standards.

MR. MARION: But, again, I would turn what you just asked into sort of a specification in the NIA and say: tell us, what is your model of, whether you go with scoping sequence or articulation, or whatever it might be, curriculum, within the states, among the states, within the consortium, and why is this model going to led success versus another model? I think that is fair to put out there.

It might be as simple as we can't get the states to agree to a common curriculum. What we could do -- I have got an idea -- how about these curricular units?

(Laughter.)

MS. WEISS: Right, but it is like,

what is your model for either being inconsistent or, if you are going to have variation, how do you support that with an assessment system that supports that variation appropriately?

MR. MARION: It is this theory of action. I don't want to beat this dead horse, but it is really is this theory of action and really specific.

MS. WEISS: Yes.

MR. BRAUN: I know I am being a cramugin here, but I want to disagree with Laurie on this notion that, just because if you look across the border and you see that state is doing a great thing, that you will naturally want to do that. I mean there's lot of evidence that all sorts of professionals can be told about a better method, and that they are not doing as good as the better method, and they say, "Yes, but I'm still going to do what I am doing."

That's true of doctors. That's true of teachers. It's true of

psychometricians probably.

But so I don't think we want to overestimate the power of evidence in practice.

(Laughter.)

MR. WISE: No, but by building a common assessment, you are removing one of the key barriers, which is it is being measured differently. It is comparable. Right.

MR. BRAUN: Yes, I think the presumption I was having is that it is not just the standards themselves. It is also the scoping sequence that you have to agree upon.

So we are a grade size lower than NEECAP is at right now. So there would be some commonality between it, and I think there has to be if you are going to do it across three or four or five, six states.

How you choose to deal with number operations in fourth grade or geometry, or how you want to articulate it, math processes, I don't know, whatever the term is, I think you can have a lot of variance.

If you have a scoping sequence across a period of multiple assessments, you are going to have to make some statement about what that assessment at time "X" is going to be evaluating, and you have to have people agree with that.

That, to me, I have been in standard development committees in math, particularly, and that is hard charge. I think you have to have it really clear in the plan, a really clear mechanism for how you are going to do that across states to get agreement.

MR. MARION: And again -- sorry, Jeff.

MR. DUECK: Well, I think there are things that you have to be very tight on. I want to go back to what Laurie started off with, and that is the whole issue of the goals, because I really agree with that. I think the goals have to be including the whole notion of accountability as being a very high flyer within that.

When it comes to the looseness, I think that, obviously, this is a pilot, and you want to see what will flourish. One of the things I would hope would come out of this is to whether or not the assessments are focused on age, stage, or grade. Because when people make a decision as to when a child should be involved in assessment is a huge issue. I would be curious to see, just from a research perspective, what would happen with using different models.

MS. WURTZEL: Okay. Just to follow up, do you all have a feel about whether we should ask each consortia to address the full range of grade spans and both English language arts and math, or whether it would be appropriate to ask consortia to differentiate based on interest and expertise, with an expectation that they might be able to somehow over time share or match up with other consortia?

MR. MARION: So like you have a high school consortia, and a K-8 maybe --

MS. WEISS: Right, and a state that was only in the high school consortium would use another consortium's elementary assessment, if they had the same standards.

(Laughter.)

MR. MARION: Good. Yes, I think you could probably get not 18 different answers from the 18 experts, but in this case I would say that I think that I would want to see somebody all in or not. If we are trying to build system coherence, I think you are all in.

MR. WISE: Or not.

MR. MARION: Or not, right. Right. That's right.

MS. WEISS: How about math versus ELA? Because that is sort of the K-12. Is that also math versus ELA, or could that be --

MR. MARION: Certainly the K-12, I guess I can't see why -- maybe you would have some reason, and I would defer to the state folks why they might say that they would be all in in ELA, but not all in math. I don't

see why --

MR. WISE: Or the other way around.

MR. MARION: Yes, or the other way around.

But I am not sure why, but we are not talking about social studies and civics here. You know, we are talking about ELA and math. So I guess I don't see why, but I would certainly require K-12 or 3-12 within a subject. I would have to think more about the other.

MR. WISE: And I would echo what Scott was saying, because if a key goal of this is to chart student's progress in getting towards readiness by the end of school, it doesn't make sense to leave out some grades.

On the other hand, I would like to see you fund more than one thing. So breaking it up into pieces is one way that you could try different things, some of which might work and some of which might not. I think maybe better would be to break it by subject than by grade, though.

MR. COOK: I have a problem with that. I mean I think literacy and numeracy are important for college readiness, not just literacy or numeracy or your model of literacy and your model of numeracy.

The idea of academic language in literacy, I mean I just think I kind of all in. If you are going to play, play; otherwise, not.

But that was the thing that came across my mind when you asked the question. If you are all in and you have a consortium, then you have limited the number of possible opportunities to be a part of the project. I would like to see as many people be a part, as many consortia have been, because I think some of them are not going to work, and some of them are going to work great.

I mean I would like to see multiple, but I have a hard time sorting out, you know, we are part of the language art consortium, but we are not a part of the math one. So we have this integrated formative

and summative assessment system working with language arts, but we are actually just doing regular old math stuff.

MS. WEISS: No, I mean I think we were thinking that you would rely on other people. It is sort of like states become vendors for each other. So one consortium would take on the ELA question, and a different consortium of states would take on the math question, and the ELA states would use the math group's output. They just wouldn't develop it.

MR. MARION: You know, we had a model like that. It was with the NRTs. You know, the vendors developed the SAT-9, the ITBS, and states purchased various ones. It doesn't seem like it would be that different than that.

I mean I think about trying to operationalize that. I think the vendors are the only one with the capacity to do that. So I think you are quickly lapsing into that model, which may not be a bad model.

MS. WEISS: Meaning that the people who weren't in a consortium would end up buying that assessment system from whoever the vendor was or partners were that were working with that state?

MR. MARION: There's going to be, within a couple of years, there is going to be five or six perfectly-aligned tests to the common core standards from all the multiple vendors, and states could be able to buy those. I mean that could be a model. I am not saying it is my model, as you know.

MR. NELLHAUS: It is interesting that you mention that because I am sitting thinking, what is the purpose, what is the rationale behind having consortia? And I have to imagine it is to have some efficiencies, to be able to do some cross-state comparisons. I don't know if it is a matter of bringing people together and getting a lot of good minds on a particular problem and coming up with more creative solutions.

I am not saying that I support this

idea, but USDE could be extending this RFP not to states, but to vendors. And the vendors would come up with the programs, and the USDE would fund those vendors whose programs they liked and wouldn't fund those that they didn't like.

And if they wanted states to really participate in these, those states would, basically, have the crux of their assessment program covered by the U.S. Department of Education.

MR. DUECK: I do like the notion of parceling things out, so that these consortia deal with different topics. I say that because, inevitably, every project needs to have a champion, a person or persons who are visionaries. They are often difficult to find in every one of the areas. So it is far better to give them the opportunity to fly a little bit faster than be constrained by the very tight constraints that this project has. So, for me, parceling out portions, and then having people come together, "I'll take yours.

You take mine," and all that kind of stuff works for me.

MS. WEISS: It does work for you?

MR. DUECK: Yes.

MR. COOK: In my view, we want to develop a system that gives us information about students, schools, districts in a state. It may be across states.

I just want to make sure, whatever system you choose and however you choose to have people participate, that we don't lose the capacity-building at the local level.

The problem that has gone on with the NRT thing that Scott mentioned, and the concern I have, is then the assessment takes on the role of, you know, the bad guy, the good guy, and this is what you've got to do, and the capacity doesn't get built.

I think you develop a system, however you choose it, in such a way that you increase the capacity of the teacher, of the principal, of the administrator to manage and implement the standards that are going to lead

to college success or career success.

That is the problem with the systems that we have. They, quite frankly, haven't done that.

MS. WEISS: Okay. We have time for a last couple of words from everybody before we take a quick break and then go to the public section.

Anyone prepare a concluding statement they would like to share?

(Laughter.)

Laurie instantly raised his hand.

(Laughter.)

MR. WISE: Good luck.

(Laughter.)

MR. DUECK: I will offer one. I referenced a little while ago the elephant in the room, and I need to let you know that you are the elephant in one sense.

(Laughter.)

The saying goes in our country that, when the elephant sneezes, we all catch cold.

And compared in size, we are only one-tenth of your population. So, therefore, what happens in your world of assessment has profound impact on me and on us.

(Laughter.)

It has been interesting for me over the last few years to have to deal with some of the fallout of things in the world of assessment that find their way into all of my periodicals within my country. So I am wishing you, from the bottom of my heart, the very best fortune.

(Laughter.)

MS. WEISS: Don't screw up, is what he is trying to say.

MR. COOK: Whatever you do, however you do it, how is it going to help kids? How is going to help teachers help kids? I mean that is the point.

I think, whatever you do, I really could care less about a really fancy-shmancy assessment system. You know, I think the ONPAR items are cooler than spit in winter,

but if they are not going to help kids, quite frankly, I don't care.

How does this help kids? How does this help teachers help kids?

MR. MARION: Yes, I think Gary's point is a good point. Then I would sort of come back to something I mentioned many times. Assessment can't be the fix to all our problems. It is a tool. Hopefully, we use it well.

I also want to say that I do appreciate what you guys are facing, and the fact that you are actually doing this publicly, airing what you don't know publicly, and then looking for us to help you publicly, instead of in a back room on Maryland Avenue somewhere, I really appreciate that. I hope the states appreciate that.

MS. WEISS: Well, thank you so much for agreeing to share your wisdom and your time with us. We really appreciate it.

Let's just give the experts a round of applause and thank them.

(Applause.)

So we are going to take just a quick, 15-minute break so we can put the podium back.

Anybody who has signed up to be a speaker, please come up to the door and see Anya. She will give you the instructions on how we are going to line you up and what you are going to do.

Everybody else, just be back in here in 15 minutes.

Thank you.

(Whereupon, the foregoing matter went off the record at 3:31 p.m. and went back on the record at 3:45 p.m.)

MS. WEISS: Thank you all.

So let me just explain how this works. The people who asked to speak ahead of time have received numbers and know what order they are coming up in, and are lined up and ready to go.

On the podium you will see lights that will tell you, it will go to yellow when

you have two minutes left, and then start blinking red when you are out of time.

So, with that, let's get going.

Everybody, please do start by introducing yourself.

MR. NEAL: Thank you.

My name is Monte Neal. I am the Executive Director of FairTest, the National Center for Fair and Open Testing.

Before the nation can successfully implement better assessment practices, it must first reject the incorrect assumptions and flawed logic of No Child Left Behind. To ensure effective education reform, including high-quality assessment, the Administration must overhaul NCLB, its draft requirements for the Race to the Top, including today's changes, and the assessment program design, to which we are responding today.

NCLB has failed to improve educational equality and equity. U.S. children have made less academic progress since NCLB came into effect than in the

preceding period, and the achievement gap has not much narrowed. Secretary Duncan's proposals to date would reinforce the errors of NCLB.

The problem is not only that tests used under NCLB are inadequate, but that the fundamental assumption behind the law has proven wrong. America cannot test and punish its way to better schools, no matter how good its standardized tests may become.

That said, the nation does need high-quality assessments that are properly used. A revised Race to the Top could provide a great stimulus for states to overhaul their assessments. This would require developing new systems of local and state formative and summative assessments that can assist student learning, help gauge students' academic progress, and provide an important source of evidence for evaluating teachers, principals, and schools.

These new systems should be built within a framework that provides flexibility

and diversity while ensuring high-quality opportunity and expectations for all students. They should start with how students learn and how best to foster that learning, building from the classroom and the school outward.

Unfortunately, the framework before us today appears designed to ensure the continuation of highly-centralized top-down state assessment systems. To a great extent, it perpetuates the flawed conceptions of NCLB. It is far too limited and will inhibit the most necessary and valuable improvements in assessment. Its structure largely reduces teachers to administering and perhaps scoring tests.

It misconstrues formative assessments, as if the issue were teachers selecting a test off the shelf, instead of responding to the emerging needs of highly-diverse learners engaged with a specific curriculum.

I would say that at least some of the speakers today extended the framework in

useful ways.

But, therefore, FairTest's first recommendation is that the assessment program design itself must be overhauled. Fortunately, there exists some well-thought-out approaches that can provide a new framework.

The Forum on Educational Accountability, an alliance of dozens of education, civil rights, religious, disability, parent, and civic organizations, that I chair, commissioned an expert panel on assessment, which included some of the people who are among your expert presenters, to develop recommendations on what a comprehensive, educationally-beneficial assessment system would look like.

The report explains how to use multiple sources of evidence, teacher evaluations of student work over time, locally-developed assessments, performance assessments of various kinds, and statewide standardized exams to determine both

achievement levels and student growth. It recommends external monitoring to ensure the quality, accuracy, and fairness of the various assessments. A system built from these elements would provide solid data for evaluating schools, districts, and states.

A growing body of evidence from the United States and other nations supports these recommendations.

Assessment is both a quantitative and a qualitative endeavor. Thus, states should be able to use these federal funds to engage in qualitative evaluation, such as an inspection system, as recommended by the Broader Boulder Agenda. Inspectors are trained experts who visit schools to observe, review data, hold discussions, and evaluate the school, and issue a report. That process is central to accountability in England and New Zealand.

Legislation introduced in Massachusetts, and supported by FairTest, would build a system that includes state

standardized test results, incorporates an inspector, and relies most heavily on assessment of student classroom work. This legislation provides the three legs on which new assessment systems should standard. The 50 percent pass-on to districts should be used primarily to design and implement local and classroom-based assessments.

In my written comments, which I have submitted, I propose concrete steps that the Department should support. These are based in large part on three attached documents from the Forum on Educational Accountability, the Broader Boulder Alliance, and the Massachusetts bill.

Some of these ideas could be incorporated into the current assessment program design that is before us today, but many would require modifying that design and going substantially beyond it in order to build new systems, not simply tinker with the current system.

Thank you.

MS. WEISS: Thank you.

Next?

MS. NOVICK: Good afternoon.

My name is Tracy O'Connell Novick. I am a former Massachusetts high school teacher. I am the mother of three daughters, the older of whom is in Worcester public schools here in Massachusetts. As of last week, I am a member-elect of the Worcester School Committee.

I want to thank you all for coming to Massachusetts. Very often, we in Massachusetts hear ourselves cited as the home of education reform, and it is very seldom that we in Massachusetts get a chance to speak our peace on education reform.

I have spent a great deal of time over the past couple of weeks reading the applicable pages of The Federal Register before this meeting, seeking to somehow answer the questions you pose regarding assessment. I have, I believe, a radical answer for you: make testing the province of the classroom

teacher.

If you, indeed, wish to have an assessment which models and supports effective teaching and student learning; which allows with disabilities and English language learners to demonstrate their knowledge and skills; which elicits complex responses; which contains varied and unpredictable item types and concept sampling; which produces reports that are relevant, actionable, timely, accurate, and displayed in ways that are clear and understanding; makes effective and appropriate use of technology; is valid, reliable, and fair, appropriately secure; has the fastest possible turnaround time, and finally, is able to be maintained, administered, scored at a cost that is sustainable over time, the only way forward is to make testing the province of the classroom teacher.

A standardized test of any kind will not meet all of these standards, but excellent teachers across the country do it

every day.

There are ways in which things would need to change. The first part might be the most difficult for some. First of all, we would need to see classroom teachers as highly-trained professionals who are, indeed, in the best position to assess their students, rather than an obstacle to be overcome or coopted in order to achieve educational excellence.

Mentoring our young teachers once they are in the classroom, making certain that we are truly supporting effective teaching and student learning, would be enormously important. It takes practice as well to make assessments that are valid, reliable, and fair.

Making assessments appropriate for all students, including those with disabilities and those who are learning English, is, again, something teachers need to master, and which can best be done with mentoring.

Master teachers do all of the above, and new teachers can learn it, but the only place to learn it is in the classroom with the cooperation of both. That takes time, and time means money. But if it is, indeed, that important, then it should be funded through your assessment grant.

Quality evaluation of teachers, a skill far too few principals are trained well in, is also important. There are those who need further training or who ought to work in another field. The time to discover this is not after they have spent years in the classroom, but very early in their careers. Appropriate training and assessment of teachers is an important piece of the student assessment, and it ought to be funded as part of assessment.

If we wish to make appropriate use of technology, we have to have that technology. Currently, too many classrooms have little, or no, technology to speak of. One cannot educate 21st century students on

Windows 95.

Assisting teachers and students in using the technology and seeing that they have the appropriate staff -- we in Worcester currently have one technology staffer for the entire city -- is a necessary part of this as well. If it is a valuable piece of assessment, it needs to be funded.

Creating varied items, assessing complex responses, making assessments applicable for a variety of students, and producing reports that are produced in a timely fashion can only happen with small enough class sizes. Having 30 children in a classroom, as we have in Worcester in 7 percent of our classrooms right now, makes this impossible.

Smaller classes means more teachers. More teachers sometimes means more classrooms. If, however, it is a valuable piece of assessment, it needs to be funded.

You might note as well that the system meets your requirements that teachers

be involved in scoring, that it is easily adaptable, that the technology involves support assessments, and is cost-effective, and that the technology used be easily adaptable.

It also goes along with truly heading toward international assessment, as this is much more like what the countries we are looking to as models are doing. Moving towards a giant, standardized assessment system or two is going in the opposite and ineffective direction.

Speaking from the perspective of 16 years of what has been called ed reform in Massachusetts, moving assessment away from the classroom and away from the teacher and the student does not reform anything. You cannot replace a teacher and a teacher's assessment by a computer or a committee.

If we truly wish to educate our children in a way that makes them good citizens who are well-educated and well-informed, we would do best to start closest to

them, in their classrooms, rather than spend $350 million to testing committees and programs.

Mentored teachers, quality technology, and smaller classes would be a great help in supporting quality education.

Thank you.

MS. WEISS: Thanks.

The next person?

MR. PHILLIP: Ladies and gentlemen, good afternoon, and thank you for this opportunity to share these ideas.

My name is Frank Phillip, and I work with student assessment programs at the Council of Chief State School Officers in Washington.

My comments address the conceptual design descriptive language of the proposed Race to the Top Assessment Program, as described in The Federal Register.

I offer these comments and suggestions to provide additional clarity and strength to the Department's proposal for RTT

assessment programs.

Current research and the bulk of today's discussion, I think, suggest that student assessment needs to be considered as a system of coherent measures that range from the curriculum embedded assessment measures teachers use in the classroom on a daily basis, the high-stakes summative accountability tests.

A comprehensive system of assessment provides a variety of information that can be used to inform the ongoing instructional process, track a student's progress, provide accountability data about the efficacy of the learning system. All of these uses for assessment information are critical for supporting student learning, the central purpose of education.

The current description as found in the document seems to be in conflict with itself or perhaps not clear in its purpose. It calls for a system of summative tests that would be used primarily for accountability, an

important function, but not the strongest approach to support teaching and learning. If the Department wants to support the instructional process that creates the learning necessary for students to raise their achievement, the RFP and guidance needs to describe a process for formative curriculum embedded assessment, interim assessment, and measure progress as well as the summative accountability test.

This balanced and coherent approach will enhance the engagement of teachers and students in the process, and thereby, emulate some of the more effective practices we see in education systems from other countries whose students achieve at high levels, such as the Albert system Jim talked about this morning.

This shift would also provide a clear target for using the local-level funds on training for teachers to become engaged and more confident in these effective classroom assessment methods.

We believe the Department should

play a significant role, leadership role, in broadening the way student assessment is understood and used to increase student achievement. We at CCSSO also stand ready to help.

We have also developed a list of 12 attributes that CCSSO believes every good assessment system should possess. They are:

The student assessment process is considered as a system with a variety of purposes such as informing learning and instruction, determining progress, and providing partial accountability information.

The assessment system addresses the depth and breadth of all standards in all areas of the curriculum, not just those that are easy or politically-expedient to measure.

The system considers and includes all students as an integral part of the design process and anticipates their particular needs.

The system considers and includes all students as an integral part of the design

process and anticipates their particular needs.

The system of assessment encourages and allows all students to demonstrate what they know and can do.

The assessment system honors the research that indicates all students learn best when given challenging content and provided with assistance, guidance, and feedback on a regular basis.

The system employs a variety of appropriate measures, instruments, and processes at the classroom level, the interim or benchmark level, and the large-scale state level.

All schools are accountable for having such a system.

Students are engaged in the assessment and learning process, and have a clear idea of how learning progresses and what they can do to improve.

Because the classroom is where the teaching/learning takes place, teachers play a

preeminent role in the assessment system. Scoring student work based on shared learning targets is common practice for teachers. New teachers and education leaders are well-educated and supported in these expectations.

New technologies constantly enhance and transform the way the assessment process is developed, delivered, and used, most notably, in providing appropriate, immediate feedback with instruments designed to support good decisions.

I have two other bullets that I will allow you to read and a handout that I do have available, if you are interested.

Thank you very much.

MS. WEISS: Thanks so much.

Next?

MS. PRITZ: Hello. I am Dr. Sandra Pritz, and I am representing NOCTI, formerly known as the National Occupational Competency Testing Institute.

NOCTI is a nonprofit assessment company serving the career technical education

field, for which general assessment is very important at the K-12 level, albeit at the high end of that level.

CTE standards and assessment are somewhat different from those in the academic content areas because of the dynamics of change in the occupational skills and because of the very large number, i.e., 70 to 100 different technical content areas in which assessments are needed.

But these challenges do not detract from the fact that the need for validated standards and valid and reliable assessments is just as great as in the academic content areas.

Furthermore, the CTE community has known and practiced for many years the fact that both knowledge and application of that knowledge, i.e., performance, must be assessed to gain a true picture of an individual's achievement.

We agree that multiple perspectives are helpful and that, as educators, we should

be caution about using one test as the only measure of competence.

Career technical education has a unique ability to bring meaning to a student's world and to keep him or her interested in the rigor and relevance of a chosen technical field. NOCTI recognizes that every technical skill contains embedded academics and that the contextual presentation of academics is often congenial to students' learning style. So the test developers mirror this concept in their technical assessment items.

We disaggregate assessment scores so that a component of the score report reflects academic achievement and helps teachers to identify academic skill, strengths, and weaknesses.

Academic reports are also one possible way to document the award of secondary credit for CTE experiences, as some states do presently.

We are anticipating incorporating the NGA national standards into the alignment

of each and every test item. We currently use the national standards of the various associations and councils in language arts, mathematics, and science to do this alignment.

We have seen many positive assessment decisions being made across the nation, and we have begun to see some cross-state collaboration and statewide articulation from secondary to post-secondary education.

Our experience with regard to online testing may be of interest with regard to your capacity that you were speaking of today. Last year 50 percent of our test orders were online; this year 67 percent.

We indicate here our willingness to help in the ongoing effort. We have demonstrated that willingness through participation in the National Research Center for Career and Technical Education and by working with OVAE and the Data Quality Institutes.

Our business requires in-depth knowledge of the process of building a test,

and we understand the need for item analysis, documented validity and reliability, and what it takes to develop a nationwide test with the proper weightings and ratings, because we use national testing.

We hope to be of service to the education community, and particularly, of course, the career technical education community in which we have the most expertise.

Thank you very much.

MS. WEISS: Thank you. Thanks.

MR. OLSON: Good afternoon.

My name is John Olson. I am here representing my two consulting companies, Olson Educational Measurement and Assessment Services and the Assessment Solutions Group.

The importance of the national educational reform initiatives for improving public education and innovations such as the common core standards now under development that could lead to states working together to develop and implement new assessment designs cannot be underestimated. However, it is

important to consider how such assessment could be developed and implemented so as to maximize their benefits to students, parents, teachers, administrators, and other citizens, while minimizing the costs of such efforts.

Thus, there are significant design options with inherent issues to be considered, and for each option, potential costs to be determined. By considering these in advance, choices can be made about the best types of cost-effective assessment system designs and procedures needed by states.

The U.S. Department of Education has set aside a large amount of money that will help support one or more consortia of states to develop common assessments that are aligned with the common core academic standards. The funds from the Department are designated to develop new assessments for state consortia.

However, a number of questions need to be considered before implementing this plan. Is this an appropriate amount of money

to spend on this endeavor? What metrics is this amount based on? NAEP? Existing state consortia data in aggregate from individual states?

What is an appropriate cost for developing the new assessments based on common standards across states? How can this work be done more efficiently and at less cost?

How can efficient services be delivered to states by testing vendors? Given that vendors will bid on the consortium work more or less sole source, what control will the consortia, much less the Department, have to avoid sole source or uncompetitive pricing?

What will the cost be to states for sustaining the new assessments in future years? How will states know if the ongoing courses will be affordable? Can states really afford the new assessment on a yearly basis?

Given the current and near-term expected financial condition of states, a new assessment must offer a significant qualitative improvement over current tests and

should cost no more to administer on an annual basis than the existing assessments, preferably less.

A well-designed and efficiently-produced assessment, combined with the scale benefits of consortia, can accomplish these objectives, but only if the details are worked out in advance.

To answer these questions, some additional planning and data-gathering would be helpful. First, a thorough review of the potential costs needs to be completed with a comprehensive cost analysis conducted that will give the Department more detailed information on the cost for each part of the new assessment, including those for startup, for implementing, and for continuing the program, with detailed cost breakdowns for all functional activities required for state assessments.

It would be wise for the Department to get all costs in advance because, based on other estimates of cost, it is quite probable

that the proposed two consortia developing tests in seven grades for math and reading and three grades for science could both do the job for under $75 million, which is significantly less than the $175 million allocated for this federally-funded activity.

Also, although states may save money on assessment development if a more varied set of assessment is used, for example, those featuring written response items, performance events, and performance tasks, states may need to spend substantially greater sums of money to administer, score, and report on these new assessments.

It is critical that the services that would be provided by vendors to the various state consortia be as efficient as possible, so that states get high-quality work without having it cost so much that it cannot be sustained by states.

While it is possible for assessment experts to provide ball park cost estimates for assessment activities, such as

development, administration, and scoring, these estimates are based on current assessment designs and single-state assessment programs. What is needed are more refined cost estimates that will roll up actual costs for multiple states into an overall cost figure.

The results will be a much more refined figure on what it will cost different sizes of states to implement their assessment designs and cost figures. The Department can be more confident as it proceeds to support consortia of states working together to create and implement state assessments.

States need to look at all aspects of using these assessments before they adopt them as part of their assessment program. Among these aspects are test design, test development, test delivery, methodology, production and manufacturing, logistics, scoring, reporting, use of accommodations, et cetera.

This information can be compared to

fair and reasonable cost for each assessment element and function.

Unfortunately, I think I am out of time. There are more comments in the paper, including more details on costing issues and recommendations made to the Department.

Thank you.

MS. WEISS: Thanks, and we have your paper.

MR. KINGSBURY: Good afternoon.

I am Gage Kingsbury, and I work for the Northwest Evaluation Association, a psychometrician by trade. I won't use any Greek letters today, promise.

NCLB has been a courageous attempt to help every child do well in school, but at its heart it suffers from the use of old testing technology and a lack of an incentive to help every child to learn more, regardless of their current achievement.

We need a system of education that encourages every child to learn as much as they can and enables every child to continue

to learn, improve, create, and innovate in school and as adults. Toward this end, we make two specific recommendations.

First, Race to the Top should require assessments that are as accurate for students who are struggling or excelling as they are for those who are performing at or near the proficiency level.

It should be the right of every student to be measured fairly by the assessment. We would not measure a student's visual acuity by measuring students near 20/20 very well and then dividing the rest of the students into just farsighted and nearsighted.

Unfortunately, that is what we do with the use of fixed form tests to measure student achievement. A fixed form test commonly provides four times as much information for a student near the proficiency cutoff level as it does for a student who is far below or far above the proficiency level. For a divergent student, the fixed form is a poor measure of what they know and can do.

Using a single form violates the right of each student to be measured well.

Second, Race to the Top should require assessment outcomes to be empirically linked to individualized recommendations for the teacher to use with each student. Tests currently used in NCLB divide students into very gross categories. This provides little information for the teacher to decide what the student would benefit from learning next.

In order to provide instructional value, the assessments need to produce timely information concerning the next steps for each student in a manner that is empirically-derived.

Students and teachers have often said that state testing is a waste of time. Until the test becomes useful for each student and teacher, they may be right.

If we plan to get to the top, we need to make sure that each and every student grows as much as they can.

Now there are very many ways to

fill these recommendations, but I will describe one test that has been proven. The use of adaptive tests with appropriate item pulls allows us to have equally or nearly equally precise measurement for every student that takes the assessment. This has been shown over the period of 35 years of research in adaptive testing by business, industry, professional organizations, the State of Oregon, and by the federal government itself.

So, by using adaptive tests, each student can take a test that is challenging and precisely aligned to the content standards of interest. Each of these adaptive tests can further provide a solid measurement of the student's achievement and a good foundation for the measurement of their growth.

Second, the use of IRT measurement scales allows us to place student achievement on the same scale that is used to measure the difficulty of the tasks on the assessment. This allows statements concerning what a student should be prepared to learn in the

classroom today, regardless of the fact that the student may be a high-performer or a low-performer. This provides the low-performing students with a pathway to proficiency, and it provides a path of challenge for high-performing students.

Now how this information is used in the classroom becomes a professional challenge for each teacher and LEA, but we need assessments that provide this precise information before the teacher can take on the challenge.

Two pieces of the approach just described have been successfully used in large-scale applications from my own organization's tests to licensure tests, to the ASMAM test used by the military.

Race to the Top provides us with an opportunity to provide students, parents, and teachers with information about the content that is challenging for the student today. We need to take advantage of the opportunity in order to make things better for the next

generation of students.

Thank you.

MS. WEISS: Thank you.

MR. MOSHER: I'm Fritz Mosher. I am research consultant to the Consortium for Policy Research in Education and the Center on Continuous Instructional Improvement at Teachers' College, Columbia University.

I found myself channeling Scott Marion today. I don't know. That happens to me a lot.

(Laughter.)

My first reaction on reading the request for input on the proposed Race to the Top Assessment System Development Program was something like what my parents would say during the Great Depression. That's the one in the thirties, not this one we have just faced.

(Laughter.)

If we had ham, we could have ham and eggs, if we had eggs.

(Laughter.)

The framework of proposed required and desirable characteristics for the summative assessments and assessment systems to be developed by the consortia of states to measure students' progress toward, and achievement of, the common standards the states will have agreed on, sets out a list of criteria, which if you take them seriously, no test publisher or assessment developer could possibly meet in the near future.

I was going to say, if they tell you they can, they are lying. Under the influence of Henry, I think we just say that they are wildly optimistic or they don't know what they are talking about.

One reason for that is simply that we don't yet know how the college and career-ready high-school-leaving standards, which themselves have not yet been finalized, let alone validated really, will be mapped back over the K-12 grades, so that the track that students are supposed to be on can be described in sufficient detail, so that

assessments could be devised to report where students are in terms of significant milestones along such a track.

One of the real virtues of this request and the framework that it lays out is that it does ask for assessments that report students' performances of terms referenced to such a track and to milestones of progress along it, rather than in terms that are referenced explicitly or implicitly only to where students stand with respect to their peers. This, I guess, is what Laurie Wise was talking about when he was talking about learning trajectories.

However, we don't really have an agreed technology for developing assessment items or exercises that could substantiate such milestones or stages of knowledge and skills or discriminate among them rigorously.

Psychometricians and publishers won't really be able to begin trying out and applying the nascent ideas they may have about such technologies until the K-12 mapping

produces grade-by-grade standards that build towards college and career readiness or otherwise defines what the track toward college and career readiness looks like, if it is not tied specifically to grade-related expectations.

This suggests that state coalitions entering into this program should expect to be involved in a lengthy and iterative, even trial-and-error, process, so both the standards and the assessments designed to measure students' progress toward them are developed and refined over time.

Don't get me wrong; I think there's nothing wrong with recognizing, accepting that attaining the goals of this program will require some time in trial and error, more of that time in trial and error than its designers may realize. If it does attain the goals, however long in time, that result would certainly justify the contemplated expenditure and more.

Still, recognition of the

limitations of our current knowledge also raises cautions about some of the other criteria listed in the framework and its requirements. We are not likely to be able to measure complex or ambitious 21st century knowledge and skills without specifying the curricular and pedagogical experiences that students should have been exposed to.

You can't have fair and valid measures of such things without such specification because you can't fairly ask the questions without knowing something about the linguistic and experiential context in which students at least should have had a chance to learn how to respond.

There are many other tensions in here, and the announcement asks for comment on the tradeoffs. I just want to say that that is a good thing. I hope you do recognize that there will be those tradeoffs and it will take time to learn how to experiment with those things and learn how to meet them.

MS. WEISS: Thank you.

MS. BALL: Thank you for the opportunity to speak today. My name is Patrice Ball, and I'm from Milwaukee, Wisconsin.

We know that no one single test, no matter how valid or reliable, is able to accurately capture all of the complexities there is to know about what students know and are able to do.

As we develop a comprehensive assessment system, I suggest that we capitalize on this opportunity to transform our assessment system, not simply tweak it. I believe we should include diagnostic, formative, and summative assessments, transforming the content and use of our current assessment practices.

First, regarding the design of the current assessment system question, in districts and in schools and in classrooms, what is measured is what matters. Currently, assessments for grade three through eight and high school in our State focus primarily on

reading and mathematics. I am hopeful that the wording "reading/language arts" will translate into "reading and language arts". What is measured matters. What is not is left behind.

I advocate for the inclusion and emphasis on writing in our assessment system. In the NCLB environment, writing has been left behind.

Learning to read and breaking the code that images on a page equals meaning is a monumental moment. However, simply decoding and even comprehending is not enough to contribute, to enhance, to make the future reality better than the current.

Historically, people were not free when they learned to read and they read. They were free when they learned to write and they wrote.

While reading is vitally important, it is simply not enough. Focusing only on reading limits individuals to a subservient role, ingesting the thoughts of others.

Learning to read and write empowers one to express thoughts, share experience, and extent knowledge, not just recirculate it.

The College Board and Advanced Placement know this will. Most subject area assessments include extended written responses. From biology to calculus, to literature, students must analyze, evaluate, justify, and explain in writing. It is not simply enough to know the content.

The way AP uses a balance of assessment items also brings excellent assessment practice to work. These next comments align to Question No. 4.

A decade ago, our District worked with Doug Rees in instituting teacher-constructed performance-based assessments in writing, mathematics, science, and art. The assessment system centered on teams of teachers who selected student exemplars. Team members used these anchor performances to train other educators. Many teachers still comment that participating in the process was

the best professional development they have ever received.

To maintain fiscal responsibility without compromising reliability in a new assessment system, we can and should use technology to assess in the assessment process. Computers, not just laptops or desktops, but cell phones and other hand-held devices can be programmed to assist and provide feedback based on examples from human anchor team members. Students thrive on feedback. The more feedback, the better the learning.

Next, regarding the types of assessment listed in Question No. 2, to be valuable tools in teaching and learning, assessments need to get as close as possible to real-world performances. Societal expectations demand that students not just know things, but are able to do things to high levels. Authentic, real-world performances need to be incorporated into the assessment system.

Rick Sigons of ATI/ETS says, "What assessment can you give today that your students would not want to miss?"

If we want students to conduct experiments, address meaningful, real-world problems, clearly articulate explanations, and use problem-solving, we must lead them and guide them in worthwhile tasks.

Take a lead from our U.S. military here. When I was a medic in training at Ft. Sam Houston, Texas, a student who earned 100 percent on his multiple choice exam was my partner for the hands-on portion of the shot exam.

Even though we had practiced usually artificial arms and read about, and were tested on the content of giving shots, when my partner pierced my skin and pressed in almost to my scapula, I knew this his multiple choice understanding was not equivalent to a real-world understanding of giving shots.

(Laughter.)

In our training, we learned about

IVs and TPR, but it wasn't until we could these things that we really were doing the work of a medic.

We should ask students to apply their skills and not just do the surface-level ways, always asking whether we are asking them to do the discipline or merely selecting those items that are easiest to assess. College and career readiness demands ability to do, not just know.

Finally, educators need high-quality, ongoing professional development to institute and maintain a comprehensive assessment system. Teachers need to understand how to construct quality performance tasks. Administrators and teachers need to learn how to provide effective feedback to enhance student performance, and all need to know how to report these to all different stakeholders.

Thank you for taking the time to take these ideas into consideration.

MS. WEISS: Thanks.

Patrice, could you just tell us, other than being from Milwaukee, who you are, what you do, other than a former medic?

MS. BALL: I am actually granted a vacation day from my employment. I am the K-12 language arts curriculum specialist from Milwaukee. I work also at Alverno College in Milwaukee. However, I am not representing those. I am here --

MS. WEISS: As an individual?

MS. BALL: Yes.

MS. WEISS: But thank you.

MS. BALL: Thank you.

MR. MARTINEAU: Hi.

MS. WEISS: Hi.

MR. MARTINEAU: I am Joseph Martineau from the State of Michigan.

I do have five points that I would like to address in the brief time that I have. That gives me five minutes per, kind of like the one minute per slide, except for minus the amount of time I just that, divided by five.

(Laughter.)

All right. So, first, the incorporation of performance measures in common assessments. I think when we are talking about a system of assessments, those performance measures are really critical, but I do want to remind the Department of the enthusiasm that there was around performance assessment in the eighties and nineties, and how it kind of fell on its face because of the problems that we came up against with that.

I don't want to get so enthusiastic about doing things that we leave behind things that have been proven to work. I am concerned that that may be something that would be a requirement in the NIA. Is that what it is?

So, really, in order for an on-demand assessment to meet the rapid reporting to inform intervention, which is what I think we are hoping for, it is important to include a majority of objectively-scored items. That does not mean they can't be authentic. It means they need to be objectively scored. We have seen some examples of items that can be

objectively scored that really do measure really high-level skills.

I think, really, that advancement in assessment is more likely to come from advancements in creating item types that can be objectively scored than in creating additional and adding more items that are going to be subjectively scored.

So I do believe that it remains important to keep some of those tasks, a lot of those tasks, at the classroom level because I believe that is where they belong because teachers and students can go back and forth with each other to really evaluate these things; whereas, in an on-demand assessment, it is one shot and you are done, and you don't really get that authentic experience of going back and forth.

So standard-setting, I am going to suggest radically, as a state testing director, that I have come to the conclusion that standard-setting is absolutely broken. Because testing has become so much more

politically-visible, the political pressures on standard-setting are really strong, and they are unpredictable. They are not always downward; they are not always upward. They are really unpredictable.

So, for example, in NAEP there's really upward pressure to create aspirational standards, and people wanted that to happen, and that is fine. In states, there sometimes is downward pressure. Oh, how many more schools did not make AYP? Or upper pressure from, say, the Chamber of Commerce or different places. And it really happens differently. You cannot predict which way it is going to go. So it is really variable.

And in addition, we do have disconnects in the level of rigor and expectations, achievement expectations, from one grade to another. Even when we have done very good work to create vertically-articulated performance standards, it still exists, particularly where we have gaps between tested grades.

So how can we address this? First, I think we can eliminate the gaps in tested grades. We've got to test. If we are going to follow, we have got to test in those grades.

Second, we do need an external criterion against which to develop performance standards. I suggest a radical departure from the way we do standard-setting. Let's just get rid of it. I suggest we just get rid of it and we say external criterion.

What is that external criterion? We look at the end game. What do we want kids to come out of school with their public education, what do we want them to come out with?

And we can actually link statistically backward, and I think people have talked about this today, we can link statistically backward. It might be just as simple as saying, what percentage of kids are college ready when they leave? Let's set the proficient bar so that it gives us the same

percentage of kids in every grade to give us a baseline.

We might reevaluate that later, but at least we can say, if they continue on this same trajectory, this is what we are going to expect. We are going to expect this kind of outcome in the end.

I do believe that we ought to eliminate traditional summative assessment. We have had some people talk about doing just benchmark. We have had some people talk about, well, maybe we ought to keep that traditional assessment, but I think they are redundant. I think that we can have some way of evaluating whether or not that is working.

I am not going to go through everything in this, but I also think that maintaining state buy-in is really, really crucial. Consortia of states are more likely to be successful and cost-effective if states are integrally involved in the activities, in the implementation of common assessments, rather than turning all the activities over to

a contractor. So the consortium is a body. The consortium is an entity. It is not an individual contractor.

I believe that states would be able to serve as really good subcontractors. Maybe one state takes the LEA. One state takes the composition. One state takes the mathematics. So that we really have this buy-in.

I think that creating a structure where it is required that the states have a really strong stake in the consortium is the only way to make this effective and sustainable.

And the last thing, conflict of interest, I am really worried about conflict of interest in the high school where the people developing the standards are also the ones who have developed all the tests. I think that also exists in K-12.

Thanks.

MS. WEISS: Thank you.

MS. SEGAL: Good afternoon.

My name is Marilyn Segal. I am the

Executive Director of Citizens for Public Schools, a Massachusetts statewide, nonprofit organization whose mission for 27 years is to promote, preserve, protect public schools and public education. We have over 70 civic, civil rights, parent, educational, religious, and labor organizations in our coalition.

Thank you for giving me this opportunity to speak today. I am excited about being here because I know that this is so important. This is the opportunity for you to make a difference, for us to make a difference, to really change the assessment system, to use that RTT money to create a comprehensive assessment system that allows students to show their strengths, rather than be held back by their weaknesses.

At no time today did anyone mention the destruction use of high-stakes testing. Our experts talked about summative assessments. But if a standardized test actually controls everything, then we stay where we are and it is not working.

In the last month, I have attended two important conferences. The keynote speaker at the first conference, organized by Citizens for Public Schools, was nationally-known educator Deborah Meier. She spoke about exciting classrooms where teachers prepare students to be informed citizens of our democracy, where students are urged to speak up and defend their position, even if it is contrary to what is being taught or commonly-agreed on. Such a classroom is a place where students are engaged and challenged.

This, unfortunately, is a far cry from the drill-and-kill, the narrow curriculum that has resulted from the national obsession with standardized tests.

Here in Massachusetts, where we are considered a national leader, we have seen too many students left behind. We no longer educate the whole child. We have cut back on the arts, music, physical education, and even recess. The one-size-fits-all approach of standardized tests that make up the State's

MCAS, especially for non-traditional learners, vocational students, youngsters with special needs, and children whose first language is not English.

Statistics show that children who live in poverty are significantly less likely to succeed in school for the middle-class children. We are failing those children. High-stakes tests have sent them away in droves.

The second conference that I attended was the Schott Foundation Opportunity to Learn Conference in Washington, D.C., last week. There we discussed the under-resourced school systems that serve our urban and rural students who come from poverty.

There is a long list of things the students need. I am going to skip ahead to what a good system will look like.

One, locally-developed and state-approved assessments to evaluate student achievement and school quality.

State-developed end-of-course exams

in English, math, science, and history that measure key content in 21st century skills.

A school quality review model to assess the effectiveness of school practices and support improvement where it is needed most.

Required annual local reporting by schools to their communities.

Accountability and intervention based on a range of quantitative and qualitative information for chronically-underperforming schools and districts.

This proposal is part of legislation in Massachusetts, and Monte Neal gave you a copy of it as part of his testimony.

It is time to look at what works. Finland has amongst the highest test scores internationally of any country. They do not have standardized tests, but they do value and trust the decision of the classroom teacher. Perhaps we should look more carefully at their model.

I asked if we are preparing students for 21st century work and life. Harvard Professor Tony Wagner wrote in a recent Ed Week commentary that he observes "only one curriculum in classrooms all over the country, test prep. As a consequence," he wrote, "many students graduate from high school today having never written a paper longer than five paragraphs, the writing format to pass state tests, and not knowing how to ask good questions, weigh evidence, reason, analyze, hypothesize, or work with others."

Business spends nearly $3 million a year teaching their employees how to write while nearly half of the students who passed the MCAS still need remediation when they go to college because they lack these skills.

MS. WEISS: Thank you.

MS. SEGAL: Thank you.

MR. DISKEY: Good afternoon.

My name is Jay Diskey. I serve as an Executive Director of the School Division

of the Association of American Publishers.

On behalf of AAP and its members, it is my pleasure to provide comments to the U.S. Department of Education regarding the proposed Race to the Top Assessment Initiative.

The AAP School Division is the principal trade association of the educational publishing industry. AAP members have many decades of experience in developing and implementing complex assessment systems in all 50 states and the nation's 15,000 school districts.

The U.S. testing industry is comprised of educators, researchers, psychometricians, and technologists with extensive experience in developing and administering technically-sound assessments.

We hope the Department recognizes the industry as a resource to be utilized as it shapes this important effort.

In my brief comments today, I would like to speak to several topics raised in the

notice.

I first want to address the topic of innovation. AAP hopes that this assessment initiative will enable greater implementation of the many innovations that the U.S. testing industry has developed. For example, over the decades, the industry has pioneered performance-based and portfolio assessments, formative and interim assessment, technology-based student assessments administered online.

In addition, the industry has developed vertical scaling and growth measures. It has developed tests that provide both normative and criterion-referenced interpretations of student performance.

The industry has developed tests of college and career readiness, assessments of English language learners, and tests for students with disabilities.

It has also created extensive and sophisticated data and reporting systems.

Finally, the industry has recently created operational best practices. These

best practices have been developed by the Association of Test Publishers, ATP, in cooperation with the Council of Chief State School Officers, with the full support of my organization, AAP.

These new operational best practices complement the standards for educational and psychological testing, which address psychometric properties of tests and the technical aspects of measurement and assessment.

Test publishers have developed these many implementations in close collaboration with the nation's states and school districts. In the case of statewide tests that are required to meet federal accountability requirements, publishers have developed the tests in direct response to state RFPs that set out detailed descriptions of what the state is seeking and what they wish to include.

Unfortunately, funding constraints often limit the scope of state assessment

systems. There's a great deal of discussion about the next generation of assessments. Much of that next generation is available now, but in most cases there has not been sustained funding for it. We hope the Department will not only fund such innovations, but we hope it will foster them through policies that are not overly prescriptive.

The second topic I am going to touch on is the capacity for continuous improvement that we hope is present in any assessment system. Over time, the testing industry has created and implemented extensive quality assurance systems. Quality assurance methods adopted by the industry include clearly-defined scoring procedures, reliable scoring technologies, ongoing training of personnel, constant oversight of the scoring process. The operational best practices mentioned earlier will further augment quality assurance measures.

The final topic I want to touch on is competition. We hope any assessment

initiative funded by the federal government allows for open competition.

The current system is a highly-competitive one where vendors are constantly updating and improving their products and services in order to remain competitive. The results of this competitive environment are innovations and lower costs. We urge the Department to encourage fair and open competition through transparent procedures and design the initiative so that no single winner takes all.

Thank you for this opportunity to provide comment.

MS. WEISS: Thank you.

MS. OLIVO: Good afternoon.

My name is Jody Olivo, and I'm a fifth grade teacher in Rhode Island. I am actually in a building that is over 100 years old. So I am kind of amazed at the amount of technology that is in this room because I have one electrical outlet, and hear of all of this wonderful technology that we hope to have and

can't feasiblely see how I would ever be able to test my 27 students in my one classroom.

I am also a proud member of the American Federation of Teachers. I come today to talk to you about the long history that we have as the American Federation of Teachers with the common state standards.

We believe that in our highly-mobile, instantly-connected world that we live in, knowledge is traveling at speeds that we possibly cannot be prepared for. Our students are going to be navigating through a world that doesn't even exist right now. They must be able to study, work, and live in states other than the one that they live in and that they were educated in, in order to be able to be successful in their own lives.

However, in the current system, where individual states develop their own standards and assessments, we as a nation have failed to develop a system that is fair to students, teachers, and schools.

The American Federation of Teachers

began advocating for common state standards back in 1983 with Albert Shanker's response to the landmark report, "A Nation at Risk".

We have been at the forefront of the standards-based movement because we see the need to ensure that our students are learning what they need to know to compete in our global society, and the need to address the intolerable achievement gap between advantaged and disadvantaged students.

In the process, however, we have learned that a conversation of common, vigorous standards is too quickly followed by a conversation of a common summative assessment. We believe in accountability, but we caution that standards and assessments are only the bookends of a true comprehensive standards-based educational system. Without the support of an aligned curriculum, professional development, and adequate teaching and learning environments, the bookends have nothing to hold together.

With standards that are detailed

and explicit, and built upon the knowledge and skills that are previously acquired as students move through our educational system, they must be firmly rooted in subject matter content and be specific enough to lead a knowledge-rich curriculum that can be mastered within a school year.

Our curriculum must provide teachers with a detailed road map for helping students reach those standards. The curriculum must focus on the content and concepts to be mastered grade by grade and include instructional resources, instructional strategies, performance indicators, and unit and lesson plans.

The assessments that we use must provide information on how well the system and/or the students are actually doing. We must have accountability for all parties. We must assist students, their parents, and the community members, and we must implement these standards with strong teaching and learning environments and school policies. It should

be us, instead of you and I.

Professional development must be aligned to all other components of the system and must help teachers and other instructional staff deliverers be able to deliver the content, differentiate that instruction, and adjust their delivery based on the data analysis. You can see that us teachers have a lot of work to do.

Time for collaboration and data analysis is my final point. The system must provide common planning time as well as individual planning time for teachers and instructional staff. This time is essential for educators to share and model lessons, review student achievement data, and discuss how to adapt instruction, planning, and assessments to meet the needs of their students.

By providing both the development and implementation of such a system, we can be informed by the teachers' collective experience and must be supported by teaching

and learning conditions that will foster our student achievement.

Thank you very much.

MS. WEISS: Thank you.

Is that it, Anya? That's it? Great.

Well, thank you so much. I am impressed that the room has stayed so full right until the end. Thank you.

I hope that you found this day as useful as we did. I noticed up at the front table here we were running out of ink in our pens and lead in our pencils, so that is always a good sign that we are learning a lot and taking a lot of notes.

Thank you very much. Thanks again to all of the experts who shared their time with us today and their wisdom with us all.

We appreciate your coming. Thank you.

(Applause.)

(Whereupon, at 4:48 p.m., the proceedings in the above-entitled matter were

adjourned.)