Colorado Department of Education

Colorado Department of Education
Colorado Content Collaboratives
Technical Steering Committee Meeting

May 10, 2012
8:45 a.m. – 4:00 p.m.
Denver, Colorado

MINUTES

Thursday, May 10

Technical Steering Committee Members
Timothy S. Brophy, Associate Professor and Assistant Dean, University of Florida
Laura Goe, Research Scientist, ETS Corporate Headquarters
Kristen Huff, Senior Fellow, Assessment, University of the State of New York Regents Research Fund
Jacqueline S. Law, Director of Assessment, Colorado Springs School District 11
Paul Nichols, Senior Associate, Center for Assessment
Guillermo Solano-Flores, Professor of Education, University of Colorado at Boulder
David Webb, Assistant Professor of Mathematics Education, University of Colorado at Boulder
Todd Morse, Associate Director, Academy District 20
Sue Bechard, Consultant, Inclusive Educational Assessment
Center for Assessment Advisors
Elena Diaz-Bilello, Associate, Center for Assessment
Scott Marion, Associate Director, Center for Assessment
Colorado Department of Education Staff
Bill Bonk, Principal Consultant, Policy and Performance
Toby King, Principal Consultant, Educator Effectiveness
Dianne Lefly, Director, Research and Evaluation
Angela Norlander, Principal Consultant, Assessment, Research and Evaluation
Jo O’Brien, Assistant Commissioner, Assessment, Research and Evaluation
Nick Ortiz, Principal Consultant, Early Childhood Initiatives

Additional Attendees
Britt Wilkenfeld, Educator Effectiveness, CDE
Joyce Barrett, Exceptional Student Services Unit, CDE
Tricia Miller, Race to the Top, CDE
Mary Pitman, Math Content Specialist, CDE
Amy Farley, Colorado Legacy Foundation
Jessica Allen, Assessment, Research and Evaluation, CDE

Bob Good, Denver Public Schools
Meg Burns (student of Guillermo Solano-Flores)

Late Arrivals
Sed Keller, San Juan BOCES
John Epps, Denver Public Schools
Krista Morrison, entered during group discussions, Adams 12 Five Star School District
Teresa Yohon, Race to the Top, CDE

Not Present
Joyce Zurowski, Assessment, Research and Evaluation, CDE

8:45 Welcome and review of agenda and meeting objectives
Jo O'Brien, CDE and Scott Marion, Center for Assessment

Scott Marion Welcomed and briefed the group on the day's agenda. We would like everyone to give a review of the assessments with a blind perspective. The elephant in the room is the resource bank, what should be in this bank? What should it look like? We can find and nominate assessments for inclusion in the bank. We also need to consider the trial and the use of the assessment banks.

Introductions around the room

Jo O'Brien We have completed Cohort I, for those of you who are advising other states, our goal is to build expertise across the state. Our goal is to have more than 200 educators in the state that will have a special knowledge of evaluations and content assessments, so that they can say "I am familiar with the design and I want to help you." Of the 10 content areas social studies, reading, writing, & communicating, visual arts, drama, music, and dance have made it through a trial run of identifying assessments. We want to share with you what we have found. We had 75 educators come together to select assessments. We really want to examine how we can show that students are learning over time.

Teachers ask us if there is one test that shows all three of these measures that Colorado will use to determine student learning:
1.) Quality criteria for one measure
2.) Multiple Measure Design Principles for Combinations of Measures
3.) Growth Measure Development
The answer is, no, these are all separate tasks.
We will use peer reviews to determine what our resource bank is going to contain. Once the assessments for consideration are identified districts may use, or not use them. .

Angela Norlander We learned a lot through this process. We may change components of Cohort II for the sake of efficiency. For Cohort I there were 4 meetings, two days each. At the first meeting, Katy Anthes presented a nice overview of SB10-191 to the collaboratives. The collaborative members used sample assessments with the first rendition of the review tool (version 1). We really are seeking to answer how one can look at assessment through a quality content lens. During the second set of meetings, the researchers visited for one day to present and discuss their findings with the collaboratives. The second meeting was spent prioritizing, and in the third meeting assessments were reviewed in depth and questions were further clarified. Once this process was completed reviewers knew what qualities we were seeking, which significantly reduced our number of potential assessment tools. During the fourth meeting, the reviews were finalized and a gap report was created, this allowed us to start seeing what areas were missing assessment tools.

Angela Norlander answered a few questions:

- Question: How many have been reviewed at this point?
- Answer: 60-200 assessments were brought to us for review initially in each content area.
- Question: How might we create some efficiency in this process? Is there a target for the number of assessments you want in the resource bank?
- Answer: Finding one assessment that works for an entire grade level is nearly impossible. We want various modes of assessment. We don't want them all to be multiple-choice assessments for instance. Having enough assessments in the bank to represent the full depth and breadth of the standards is the goal, not necessarily a target number of assessments. One approach we took towards being efficient was that everything was reviewed in small groups in each collaborative. If the members were reviewing individual items, such as released NAEP items, they would review them as a cluster instead of reviewing the items independently.

Jo O'Brien We received about 400 assessments; they had a lot of variability. (Angela handed out a few documents). Researchers stated that the original template was cumbersome. The CDE PO Exhibit A Statement of Work page two, states "The vendor shall identify and evaluate acceptable academic growth measures in preschool through twelve grade Social Studies….etc" we wanted details on the format/mode, we wanted to know about accommodations, and many of our samples did not have these. For example, with music we had a lot of samples, but some assessments were expensive, and we did not want to create unreasonable expectations. We wanted to make sure the assessments met our standards, and that they could be implemented.

Timothy S. Brophy "How were teachers asked to manage and be aware of their biases?

Jo O'Brien We had professional facilitators coach teachers on evaluating their own judgments and biases. It was not okay for teachers to just say no, they needed to run things through a filter before giving us their feedback.

Sue Bechard Who determined the alignment issue? Was there an agreed upon set of criteria, especially for getting agreement on a cognitive design?

Jo O'Brien The collaborative members reviewing the assessment determined the degree to which the assessment was aligned to the standards using the questions from the content review tool. The members received training on DOK and used the DOK of the standards to decisions about whether or not an assessment had the appropriate cognitive rigor to be sufficiently aligned to the standards.

Scott Marion When we ask for technical documentation we automatically favor some tools over others. Many new assessments and performance-based assessments do not have this criteria but that doesn't mean the assessments shouldn't be used..

Jo O'Brien We were very pleased with what we got from our researchers. We set a minimum level of requirements, but these can always be improved. Researchers did not record DOKs. They only had 3months to work on this and were very thoughtful about what they were reviewing.

Laura Goe Are these the right things that we are asking of the experts on page 2? I imagine that we are looking at the quality measures. We need to look at assessments that can measure growth overtime, and what we ultimately want is this piece, but we need to clarify that we are trying to measure student knowledge at a particular time and not over the long run. This is very unclear in the scope of work.

Jo O'Brien There were confusions over this throughout the process, and we did have to clarify this later.

Scott Marion I'm not sure if the people are clear on the difference of measuring student learning in terms of the Colorado academic requirements in a point of time.

Laura Goe We are looking at a point in time, but this is unclear in the scope of work.

Kristen Huff This scope of work will be revised for Cohort II. I think with the CDE Qualified Assessments for Measuring Student growth we are making this very difficult. We maybe need to think of buckets, was this developed by a commercial vendor or a smaller vendor, are we preferring one over another? We will need a way to ensure that we are not.

Sue Bechard We may want them to look at the purpose that the tool was created to evaluate.

Guillermo Solano-Flores We may want a sampling matrix for how it is going to be used in an evaluator system, one that looks at content and skills and measures whether the assessment task is good and measures where it is going.

Scott Marion We will come back to this later. We will have a blueprint, but we need to ask if we are we trying to hit every cell in the matrix, probably not. We need to think about this question. Are we sampling the domain appropriately? CDE could probably help us with this, but Districts could still poll in one particular cell very poorly. I would be happy if we had an extensive search that wasn't limited to formal technical documentation. I think the Content Collaborative Members are getting very well trained. I don't want to cut out good assessments.

Paul Nichols Relaxed creates an impression of not satisfying certain technical expectations. But validity and reliability are necessary components. We need informed judges that make holistic decisions on assessments that will provide valid and reliable assessments.

Jo O'Brien Maybe we should review some of our selections because some of this selection process is risky. We may find that it may not be the usual suspects.

Kristen Huff We need to emphasis that this is a use at your own risk process. Do you aspire to make any kind of statement that these are model assessments to be used to create a teacher's own assessments?

Jo O'Brien Yes, we want to provide examples that can be adapted. We want to be capable of stating that these assessments will meet the minimum criteria that we are looking for in an assessment.

Guillermo Solano-Flores When you find a task or assessment that has great quality and is sensitive to performance differences, then you may want to make a blueprint of the assessment structure from the assessment. It is possible to create templates that would allow others to create an assessment that meets a basic level of requirements.

Jo O'Brien We want to feature examples that meet our criteria as well as high quality examples. We can meta tag these so that they exist, but the body of evidence should be balanced.

Scott Marion Task shells are a great idea. It gives people a solid starting point.

Sue Bechard I wonder if the researchers could give some indication of a holistic judgment about a level of flexibility but also a level of standardization.

Jo O'Brien Introduced the high quality assessment content validity review tool. The questions we asked were that people ask themselves 'would you want this used to judge you? And, 'Is it aligned to CO Academic Standards? Is it aligned because it can be scored using clear guidelines and criteria? Is it fair and unbiased? Are there increased opportunities to learn?' These were the questions we wanted the collaborative members to work with as they reviewed an assessment.

Laura Goe There should also be something that is valuable to the teacher? Sometimes a project demonstrates what a student learned more than a multiple-choice test.

Angela Norlander We were trying to get educators to look at a variety of assessment tools and see what was actually useful.

Jo O'Brien We wanted them to pick up the assessment from the options, and see if it met the criteria outlined in the review tool. We found a way to visually show case what the reviewers were saying.

Bill Bonk Was everyone looking at the exact same thing?

Jo O'Brien Yes, we looked at grade level, standards, and grade level expectations. This allowed the process to be far more granular and allowed us to look at a performance tasks. It was absolutely necessary for everyone involved in this process to know the standards in order to do the reviews, but we also had the standards available for everyone to review as they walked through an assessment.

10:17am Review of revised version of content review tool and accompanying summary
Dianne Lefly, CDE

Dianne Lefly This started with a huge document that was nine pages long, reviewers could check something and then make a comment. They could choose if an assessment was in complete agreement, if it was sometimes in agreement, or if it just was not in agreement. I created numbers to represent the responses of the evaluators. There were a number of topic areas and the reviewers scored each of them. The numbers allowed me to create percentages in order to evaluate the assessments across content areas. The total scoring that you can see on the PowerPoint, deals with the scoring criteria, this can be found on page 3. If someone marked an assessment as a one, two, or three, then they had to explain why they selected that score. This is a simple summary.

Jo O'Brien People were not allowed to see the assessment scores until after they were finished reviewing a number of assessments. The reviewers would number assessments briefly, and this would allow them to select a few visuals to predetermine a quality assessment tool.

Scott Marion Why are total scores low? Before we state whether there is a problem we need to evaluate whether the rubric or scoring tool are producing the lower scores.

Kristen Huff I was happy to see the S. Africa scores lower because I feel like we should see more variance than what is presented on the Content Review Tool Summary: Scored Social Studies Assessments.

Jo O'Brien Items in the bank are linked to the reviews, so that you can see qualitative and quantitative scoring.

Dianne Lefly Refers to the 'Content Review Tool Summary: Scored Social Studies Assessments by Criteria' this added another level of depth to the reviews. Then I created yellow boxes that would score the assessment.

Bill Bonk Are you guys covering what worked and what did not?

Dianne Lefly Yes, they state this covers standard one but not standard two for instance, and the evaluators evaluate the assessments constantly questioning if they would want this assessment used on them.

Scott Marion I'm not sure that the percentages are the right representation for this data. Kristen's point on variance is important for us to consider as we move forward on this.

Elena Diaz-Bilello We also need to think of the fact that some criteria weigh in a little heavier than others.

Jo O'Brien Let's look at the inventory. This is a screen shot of what we have captured so far. The assessment in blue is important for us to track so that we can look at this later on if someone asks us about it. You can sort these by assessment type, grade level, mode, standards, and depth of knowledge. You can also see whether or not others recommended them. The name of the assessment acts as a hyperlink, so that one may review each assessment through this site.

John Epps We do have some gaps of information in our matrix. Recommendations were only given if an assessment met the current state standards.

Jo O'Brien Introduced the slides that indicated the number of assessments reviewed by category and grade level. For instance Music, Elementary School, had 34 assessments total, blue items were not reviewed by the end of 8 days, the dark blue were reviewed, and the yellow were partially or fully recommended. A number of these slides were presented.

Jo O'Brien We will be looking for grade level expectations. This table will be hyperlinked. The goal will be to have multiple assessments and expectations. Over time, we will improve the bank by retiring some assessments in lieu of newer, more precise assessments.

Scott Marion Technical steering committees are always good at creating work; if we had a matrix with the standards we could also have depth of knowledge levels. We can have an assessment that indicates which marks it meets.

Guillermo Solano-Flores Do the yellow cells have one or more than one assessment?

Jo O'Brien Some do and some do not.

Guillermo Solano-Flores We should have a hybrid of this that shows multiple assessments and requirements.

Jo O'Brien For the assessments that are in here, and are recommended or partially recommended, it is nice to see that they show grade level and content requirements.

Sue Bechard This could be useful for creating continuity.

Jo O'Brien Each of these show grade level expectations, by topic, and basic requirements, this will grow over the next few years.

10:57am Break

11:12am Reconvene, Blind Review of Sample Assessments in Groups:
Angela Norlander, CDE and Elena Diaz-Bilello, Center for Assessment

People were broken up into the four groups:

Elena Diaz-Bilello Group 1
Jo O'Brien Group 2
Dianne Lefly Group 3
Scott Marion Group 4

Scott Marion gave a brief overview; we are most interested in your opinions briefly on the quality of the assessments but more interested in the quality of the ratings themselves.

Group 1 We had reading and writing and were in agreement that the four tasks with rubrics of the tasks were impressive examples of assessment. We looked at reading, writing, social studies, drama, and dance. We started with one that made us cautious moving forward. The rubrics need more work in the interpretation side; it needs to be clearer with examples. As for ratings, we felt that dance was out of our field, and perhaps that was why we didn't understand it so well. This example had good tasks that were impressive.

Todd Morse There was a spectrum that allowed reviewers to look at reading and writing and quickly make a decision on whether the assessment would be useful, but on some of the others it would have been harder to decide if the assessment tool would be useful.

Group 2
Sue Bechard We did an intensive review of one assessment that was interesting as it was presented, but we were wondering how the reviewers could really use this as an evaluation method. It seemed like a teacher, without very clear technical measurements, could arbitrarily select standards. This was the Wyoming tool, we liked it, but it lacked specific technical elements. It would be helpful to have a cover page that gave the researchers more information about each evaluation tool. This would at least advise them on what they were preparing to use. We only got this information from an outside source. Without Krista it would have made no sense. We need something that has information on the criteria before we use these assessments. We need to be measuring kids based on the targets we are trying to reach. The scoring rubric had some examples, but it lacked sample text of this is one way a student might respond on a given text. The inter rater reliability might be really low if we don't have specific examples. We want to give the evaluators the chance to have clear examples of how answers could be rated.

Guillermo Solano-Flores See the absence of these makes it hard for us to have inter rater reliability, and this is a hard target to meet already.

Sue Bechard On the fair and unbiased page we were not sure how the raters would be able to evaluate 3a. What do we need to tell teachers in order to ensure that they develop clear materials? We had comments about the accommodations piece. The example we looked at did not have information about accommodations. Should these be included? Should IEPs and targets be included? This list that we are using may not be an appropriate method for us to apply to the assessments.

Guillermo Solano-Flores Even if accommodations were clear to the evaluator, it still would be unclear if they were provided adequately or appropriately. The wording of the tool does not provide specific accommodations given the students special needs either. Some of the evaluative criteria allow for a wide range of interpretation, but the State requirements are often clear. The answers for these also change when educators put themselves in the student's shoes as they review their assessment.

Scott Marion This brings us to a great question, I wonder if each assessment that goes into the bank should have specific accommodations with each assessment, or if there should be a standard set of accommodations that is used for each assessment?

Sue Bechard Maybe you use a basic set of accommodations for each GLE. For math we know that you do not allow calculators, but maybe there are specific issues to be considered depending on the topic.

Guillermo Solano-Flores Whatever you decide to include in this rubric, you will have to decide on a certain set of accommodations. In general, this information will probably not be stated in each evaluation assessment tool, and even if it is stated, it may not really address the issue of fairness.

Toby King Perhaps this is a step in the process "does the student have an IEP?" and then we use that to determine the accommodations.

Scott Marion Is it too much of a burden to address the issue of assessments with accommodations? Teachers have a vested interest in having their students look good.

Toby King This really is about more than evaluation.

Guillermo Solano-Flores You may end up with 25 tasks, but for now you should start developing your own accommodations to determine what is allowable. The best set of accommodations so far is NAEP. We have to distinguish between reviewing the task from different perspectives: teacher, auditor, etc. For instance, the auditor will be looking at the assessment in terms of its value and inter rater reliability.

Group 3
Dianne Lefly We looked at all of the assessments and noticed that the evaluations were probably loose. We noticed that teachers might not know what defined a good rubric. We don't want everyone getting the same score and being happy when our goal is really to measure effectiveness. We should train people about what makes a good rubric.

Jo O'Brien We want to grow many educators in the content collaboratives. We need to transmit information through a variety of channels. We had people asking what is "fair"? We could use videos to show footage of expert descriptions and shared terms. Rubrics are difficult for people to comprehend.

Dianne Lefly We noticed that evaluators would say that rubrics were weak, but they still gave it full points.

Laura Goe We also noticed that some tools had high scores and no comments, while other tools had lots of comments and lower scores, but they actually had better content. We were wondering if the higher scores were just higher because they had nothing to really look at and evaluate.

Bill Bonk We may want to have testing experts review the scoring rubric.

Dianne Lefly If we had a rubric that you used over the course of one year, then it might actually measure the student's growth overtime.

Laura Goe We should create a rubric first for each task.

Scott Marion We may get better evaluations from rubrics created for specific tasks. Perhaps we need a task rubric shell?

Guillermo Solano-Flores We always have to adjust tasks and rubrics overtime because rubrics are only mediocre on their own.

Sue Bechard What are we finding out from Cohort I and will there eventually be a way to get reliability?

Jo O'Brien We need to be able to re-evaluate some of these as we go through the process. Furthermore, we shouldn't make an assumption that every teacher knows DOK. So it is a great idea to have professional development and growing DOK knowledge.

Krista Morrison We had some that we looked at as a set, and we had other ones that we really evaluated with a close eye. Many of the rubrics are not very good, and this made it harder for many of us to score various assessments. So we do need to educate evaluators on what makes a good rubric.

Laura Goe We need a rubric and a scoring tool. If we combine the two of these then we will get more reliable data.

Toby King The pilot will be a really important step in this process.

Todd Morse We need to inform end users of strengths and deficiencies in the rubrics. The review tools make it easier to do the evaluations. If we use one that is already created it saves us some time.

Scott Marion If we are able to demonstrate the strengths and weaknesses of the rubrics then we build stronger evaluations and inter rater reliability.

Group 4

Kristen Huff We focused, as Group 3 did, on the rubric issue. We reviewed drama, visual arts, and dance. This was difficult for many of us. The weakness is the rubric and we agreed with the evaluators but we thought they were too lenient. A task that fulfills high marks on alignment would only be accurate if the rubric was stringent. We need more in terms of the grain size, perhaps a seven point liker scale. Independent grading could be used and then averaged.

Dianne Lefly If there are too many points on a scale, then people select the middle more often.

Kristen Huff Consensus can take a lot of time, so an average maybe useful. Perhaps there could be some options were you could force some choices. We also had a great discussion about accommodations, but this is not an area that I am familiar with providing.

Scott Marion We thought content knowledge was important in terms of doing evaluations. We reviewed a few tasks, one in drama where they created scenes for nursery rhythms.   We looked at how the evaluators scored some segments of this tool, and while we agreed with elements of their review, it was hard to reconcile the inherent flaws of the rubric and vague descriptions of tasks.

Kristen Huff I got lost in the pages and pages of materials; it was hard to find the segment of the assessment. I think this needs to be simplified.

Angela Norlander We tried to have these in a set format. If we removed elements of the tools to make them more user-friendly then it lost some of the important elements of the assessment.

Laura Goe Maybe we want to start building capacity?

Jo O'Brien The first phase of Cohort II will be to start building the bank, and eventually we do need to build capacity, but we cannot do it without creating a basic set of tools in the bank.

Scott Marion We have to think about our theory of action. We have to change things at the local level before we can really move forward. We did feel like the reviewers did a pretty good job, and it sounds like other people at the table agree.

Elena Diaz-Bilello What about the audience issue, if you are a principal you will need more information. If you look at dance you need specific information about tasks and reviewing methods.

Timothy S. Brophy What is the extent that these are being used for cultural biases? Nursery rhythms assume a cultural understanding that students may not have, so we need to make sure that cultural biases are addressed in our assessments.

Laura Goe Some cultural issues may be required in our standards.

2:20pm Structuring the Resource Bank

Scott Marion The resource bank was originally created to allow people to share assessments and to support tested and non-tested subjects and grade levels. We have to think a little more about the resource bank. We could have a controlled item bank where we indicate that you must pick one from each category say A-E. We also need to know when the bank is full, and we need to make sure the same assessment is not being used over and over again because clearly students would become better overtime.

Jo O'Brien We need to think about how the bank will look, for instance should it be organized around a standard? I would like you all to tell us about user-friendly designs. We need to think about weaknesses that could be created. We need to think about the look and feel that we seek in our resource bank.

Paul Nichols Have we thought about equating, so that we have comparable results overtime?

Jo O'Brien We would need some kind of algorithm eventually to show growth overtime.

Toby King When we talk about standards we have them for students, educators, etc. So what do we mean by standards, and can they be used on various levels (Elena said this should be clear). Is this clear to everyone? It should be.

Sue Bechard Are we thinking about a comprehensive blueprint?

Scott Marion: I used the term blueprint in hopes that people would understand that we are looking for instruments and modules.

Jo O'Brien A blueprint is a loaded word, and maybe that isn't quiet what we are trying to do here. We really want to be able to take an inventory of assessments so that people have options. What look and feel might the bank have? Think about when you go onto Amazon, and you are looking for books and movies, is this kind of look and feel we want to mimic for our resource bank?

Laura Goe Amazon is a great example, but we should also be able to compare items. Say you wanted to look for English standards, maybe we would craft it so that you could look at the measures with additional considerations, and maybe we should even have a review feature for the assessment tool. This would allow people to have a user-friendly experience. Maybe we could have someone from the DOE categorize the assessments in an A, B, and C method so that an educator could combine assessments to get a richer evaluation.

Kristen Huff These assessments will be used to measure student academic growth, but there are other buckets that look at issues of a student's individual growth, so we need to categorize these.

Toby King We could Meta tag assessments in terms of topics and grade level; maybe they are appropriate for measuring a group of students, but they may not be useful for measuring individual growth, so we could indicate this with a Meta tag.

Scott Marion I don't see the difference in assessments that would be useful at the group level and not at the individual level. How much is 'shared?' I guess this is really tomorrow's discussion. Should the usage of an assessment go beyond a given topic and grade level, or is this really going too far?

Bill Bonk We have a constrained group of users. We can work with a vendor to see how much it will cost to get a particular interface, and we can decide to have some elements of an interface later on.

Paul Nichols The methodology you use to make the interface user friendly, like Amazon, should meet your overall goals.

Jo O'Brien How do we make sure we are user friendly and not mediocre?

Sue Bechard We need something that tells an evaluator that they must use multiple evaluation methods.

Jo O'Brien We have a user interface question versus the wizard approach to a database. We also want to ensure that reviews are not emotionally based.

David Webb We could have users in the database clarified, say expert opinions verses teacher's opinions.

Jo O'Brien What if we have some filters in the resource bank? We do not need a twitter. We want to have a simple summary.

Toby King If I want to use the Wyoming example, then I need to know about it, and I need to have approval from those above me to use it. The 1338 council could be a gatekeeper for this process.

Scott Marion We do not want teachers picking evaluations that will make them look good. We need evaluations that are useful.

Jo O'Brien How do we make the resource bank one that constantly grows human capital? It should not allow an educator to constantly select lower level evaluations.

Scott Marion We also need to think about SB191, how are teachers in non-testing grades/classes going to be evaluated?

Laura Goe The items that end up in the resource bank will meet a certain level of standardization, so this should eliminate the chances of a teacher selecting less effective evaluation assessments.

Paul Nichols What is it that the system hopes to accomplish? What are our goals? Once we get these established it will be easier to decide what we have in the resource bank and how it is set up.

David Webb Currently, resource banks do not have reviews and ratings; we see these more for entertainment items.

Laura Goe We usually have a lot of information on items we use before we use them in the field, but these assessment tools are new, so I think the reviews could be a little more useful.

Jo O'Brien When I was out of town I needed a doctor and I found a website that had really useful information about various doctors in the area where I was. It gave very detailed information that was linked to the doctor's abilities. It didn't allow me to look at parking in the area, but it showed surgery success rates and other useful information.

Process observation-we have gotten so much out of this so far, and I think we should take this conversation to the point we need to get to in order to address our needs.

Scott Marion Consumer ratings, so far we have a lack of consensus regarding their usefulness.

David Webb We have to be careful if we allow users to rate a resource because of the political implications that could be involved.

Jo O'Brien Yes, we don't want to see an assessment tool negatively rated by a few users and then that sets the precedent for that assessment, when maybe the reviews are negative and really it was a very useful tool.

Laura Goe I like to think that teachers would leave comments in order to improve the usefulness of the evaluations.

Kristen Huff Why do we need reviews in the element data bank?

Timothy S. Brophy We need to make a distinction between the usefulness of the reviews and what kind of issues we will allow users to rate in the system.

Scott Marion Maybe there is a way to enter tips on use for fourth graders, etc.

Laura Goe I feel like the users could add to the usefulness of the assessments by adding valuable feedback on the assessment tools.

Jo O'Brien It would be interesting to see who is saying what about which assessment tools. Do we want to distinguish the roll of the person giving the feedback?

Laura Goe Yes, I think this would be really useful.

Toby King This could help districts with improvement processes and planning as well.

Scott Marion TCAP is not the same assessment every year, and it would be nice to make sure that the same assessment is not used over and over again every year. If a teacher finds an assessment that they like, we still want them to select a new one for the following year.

Teresa Yohon You could add a bank of administrative questions.

Jo O'Brien We want to think about how the bank will look and feel, but also how it will function, especially if it had features like a wizard? Perhaps we would allow a person to override the wizard if they need certain tools.

Scott Marion What is the requirement for the student performance in growth element? Are we looking at the non-tested subjects? Different states are struggling with this component. We don't have to use student growth objectives. If something works well in seventh grade it may also be useful later on in another form.

David Webb If users login then you can get a mapping of things that teachers use overtime.

Mary Pittman We are really trying to steer clear of power standards, especially in GLEs. The assessment framework is the standards. We need to be careful when we talk about these with the public.

Scott Marion I think this is a language issue.

Jo O'Brien I like the idea of logins and mapping what a user is doing.

Toby King Is it possible that this bank could be open during a window of time since districts are supposed to follow a timeline and teachers are supposed to have these approved?

Laura Goe Why would we want to limit the access to these?

Toby King So that teachers and districts are not using the same ones over and over again.

Timothy S. Brophy If it is an effective measure of evaluation then we don't want to remove it.

Jo O'Brien We should age the assessments so that we can track them overtime and rotate them through the bank. We should have the newer assessments at the top. Typically we do expire assessments.

Timothy S. Brophy We need the bank to be sustainable, so they should be periodically updated and revised.

David Webb We could also alert users to the fact that a newer version exists of an assessment.

Kristen Huff If I have already downloaded an assessment, then there is nothing to keep me from using it every year, but I could also download it and never use it.

Paul Nichols How are scores collected and aggregated? Shouldn't we be able to see which assessments are actually used?

Jo O'Brien The person who invented the trip advisor for Delta didn't understand how to get to different locations, and we need to think about things that we would like to see our engineers think about as they create our resource bank.

Laura Goe My husband created a device that allows me to see how many people have looked at a document that I access from the Internet and this might be useful for us in our resource bank. It might also be interesting to see where people are located and what people are looking for on the database.

Scott Marion User interface software designers would be good for us to speak with before we set out on this course.

Jo O'Brien We want to be stewards of this database and we want to make sure that people are not just cutting and pasting parts of evaluations together in a haphazard manner.

Kristen Huff There are a lot of conversations taking place right now about tagging information within a database.

Teresa Yohon We are part of this conversation.

Jo O'Brien If a teacher wants to see resources that can assist them with meeting higher standards in evaluations we want them to be able to access these on our site.

Scott Marion There have been other organizations seeking to do this as well, so we should see what is currently out there.

Kristen Huff Learning maps look very similar to what Wireless Generation is doing.

Laura Goe Ideally it is possible to have a sample test within this engine.

Paul Nichols It might be nice to include how long does it take to administer?

Jo O'Brien So if we think pie in the sky, it would be nice to eventually have small engaging commentaries that might make it easier for say a fourth grade teacher. It could direct them to new methods or university classes that help them achieve their professional or evaluation goals.

Laura Goe We could use other features of Amazon, such as listing recommendations, and showing users who looked at this also looked at_________.

Jo O'Brien Thank you all for assisting us with thinking about how we want this to look.

Jo O'Brien Angela, thank you for getting those gaps clarified for us in the matrix, without it we wouldn't be here right now. Then Jo gave a brief overview of tomorrow's Agenda.

4:15 Wrap-up and adjourned for the day

Friday, May 11

Technical Steering Committee Members
Timothy S. Brophy, Associate Professor and Assistant Dean, University of Florida
Laura Goe, Research Scientist, ETS Corporate Headquarters
Kristen Huff, Senior Fellow, Assessment, University of the State of New York Regents Research Fund
Jacqueline S. Law, Director of Assessment, Colorado Springs School District 11
Paul Nichols, Senior Associate, Center for Assessment
Guillermo Solano-Flores, Professor of Education, University of Colorado at Boulder
David Webb, Assistant Professor of Mathematics Education, University of Colorado at Boulder
Todd Morse, Associate Director, Academy District 20
Sue Bechard, Consultant, Inclusive Educational Assessment
Center for Assessment Advisors
Elena Diaz-Bilello, Associate, Center for Assessment
Scott Marion, Associate Director, Center for Assessment
Colorado Department of Education Staff
Bill Bonk, Principal Consultant, Policy and Performance
Toby King, Principal Consultant, Educator Effectiveness
Dianne Lefly, Director, Research and Evaluation
Angela Norlander, Principal Consultant, Assessment, Research and Evaluation
Jo O'Brien, Assistant Commissioner, Assessment, Research and Evaluation
Nick Ortiz, Principal Consultant, Early Childhood Initiatives
Joyce Zurkowski, Assessment, Research and Evaluation

Additional Attendees
Britt Wilkenfeld, Educator Effectiveness, CDE
Mary Pitman, Math Content Specialist, CDE
Amy Farley, Colorado Legacy Foundation
Bob Good, Denver Public Schools
Meg Burns (student of Guillermo Solano-Flores)
Sed Keller, San Juan BOCES

Late Arrivals
Tricia Miller, Race to the Top, CDE

8:30 Considerations of use context—particularly for determining student "growth"—in the evaluation of assessment technical quality
Scott Marion, Center for Assessment

Scott Marion Introduction. Yesterday was a great discussion and I think we learned a lot from it. I think it really validated the work of the reviewers and the way that they were evaluating assessments. We do have some concerns about accessibility and the rubrics. We need to have training with the collaborative members to build on the general expertise in this area. I think we saw both good and bad examples yesterday. In terms of structuring the bank I think we got some things for CDE to work on and we got an idea about how the bank will be structured. Would anyone like to share anything this morning before we start? No, okay. Why don't we start with this student growth piece? We are about to see stuff that Toby King deals with everyday.

Toby King (A general note to see the Content Collaborates Handbook). Different districts can weight the first five of our Teacher Quality Standards differently. Today we will be focusing on student growth and the other measures aligned with CDE guidelines. See the SCEE Final Report, page 10. .

Scott Marion If you don't have a test to go off of you can construct a mini-VAM to get a measure for the student.

Toby King Evaluate the technical merits of calculating growth. To get a good measure of growth you need a lot of students, and a district likely doesn't have enough students to get a good measure.

We should consider using student growth objectives or other goal setting approaches. We may weight a TCAP score with some of these other tests in order to measure growth. In our state board rules we define student growth one way, and when people in CO think of student growth they always think of the student growth model. So how can we use the student growth model to get an accurate model that will actually test student growth? We cannot test 10,000, or even 1,000, students to get an accurate measure of growth. Student growth objectives are sometimes rigorous and sometimes relaxed. Even if we had complete confidence in the assessments, will they be used and applied as they should be used and applied? For my growth, I can see my students and evaluate them daily, so that when I evaluate them using the assessments it is done in a useful manner. How do you combine the growth with the professional practice? When we apply the matrix that we are supposed to use, then the scores may or may not be an accurate gage of the teacher's effectiveness.

Scott Marion We cannot look at everything on the practice side, but we can look at how we can think about the inclusion or non-inclusion of the resource bank.

Jo O'Brien Toby is talking about the larger gestalt about concern and need. So if I can, I want to pose a question, rather than just talk about how do you combine high quality assessments, Colorado needs your advice on whether the state model, which allows districts to elect to use different models, (value added models, ups and downs of SLOs, etc….) can begin to determine where other states have gone out ahead and landed on a decision of a model to use. We want to see what we can learn from other states.

David Webb Can we add to that what are the pros and cons of the current growth model that Colorado is using? As a current parent, I see growth charts for my kids and I find it to be useful.

Scott Marion So far Colorado has received positive feedback on these. We have a terrific information system with a great information package. Colorado has been an innovator. It takes a while to roll out a good model, and many states want to do this in a few months or a year, this is not possible. CDE is working on validating the student teacher links.

Toby King Can you repeat the question David Webb?

David Webb I want other states to look at Colorado's model since CSAPs growth is going to be integrated in the new model.

Scott Marion We are also going to be looking at classes and grades that haven't been tested in the past.

Jo O'Brien This question is under every educators skin, the growth model is intuitive and attractive model because not all kids start at 4th grade plus one month. I think the accountability system makes since, and some districts want a more granular way of measuring growth overtime, and this is not what Colorado has supported so far. We are doing something with the current model and bridging it to measure a more granular approach.

Elena Diaz-Bilello No one is saying that the current method is perfect. Some districts see the current model as being erratic or up and down.

Scott Marion Is anyone here going deeper on this?

Toby King Harrison School District 2 has decided to put their effort into constructing peer-reviewed assessments that teachers can administer to students in their classes. First, they run copies of the assessments without names and double score them and then aggregate the scores. For growth they allow the educators to pick a keep up goal or a catch up goal and they write goals similar to an aggregated 45-55 score. If the students have a score of 38 the teacher may make a grow goal of 45, which might be attainable, but if they were to select a goal of 55, then this may not be attainable in the given time period. What does the machine look like that aggregates student growth? Harrison has made a large number of these, but many districts do not have the capacity to mimic Harrison.

Bob Good Returning to your question earlier Scott, yes with cautions. Furthermore, trying to have goals on student percentiles is like trying to predict the weather. Teachers can go from 20-80 and that is very common.

Sue Bechard When you are talking about growth, what is the assumption on the measure of growth?

Toby King Reads the decision of growth 1.22 state board rules to define growth.

Sue Bechard What does change in achievement mean?

Scott Marion That is part of the general question states are looking at.

Laura Goe You must use rigorous comparable measures. Many states that have implemented this have legislation that is requiring the new measures. Most states only have a few tested grade levels and subjects though. The Delaware Model for Race to the Top was hard work and it didn't get results they hoped for. They were trying to figure out if they already had tools to measure growth within two points of time. There are plenty of ways to measure within a single point of time, but not many with pre-and post measures. There is now a consultant there that is trying to provide more training and oversight in developing measures. School wide value added is being used there now, and Tennessee is also doing this, but this model has deficiencies. The idea of creating comparability should allow a common set of measures for all teachers. Student learning objectives started with incentive pay and some districts in Colorado tried this approach. Teachers would create objectives with the supervision of their principal. The end result has been widely adopted because it is simple. The other option is to try to find a way to measure every grade and subject, for instance Hillsborough, Florida, got a lot of money to do one of these models, but it is costly and time consuming. These are not intended to be all multiple choice; projects would likely show better measures.

Bob Good In Hillsborough, there are 720 tests with 720 models, these are all multiple choice, and for revision the testers run analysis and send a list of items that need to be addressed but these lists are vague. New items replace the old items, but they usually are not pretested. These have been in existence since the mid 80s. It is just a resource. The staff for this is small, but they did get Gates money. The performance model is dangerous.

Laura Goe Yes, how do we measure welding, music, etc? These courses require a performance.

Scott Marion Some will say that multiple choice doesn't measure anything.

Laura Goe In Rhode Island, student-learning objectives allow teachers to calculate a percentage that they hope to have proficient by the end of the year. Teachers create their own targets, but they also work with their administrations on this. My big problem with this is that teachers probably pick lower percentages they could meet just to make themselves look better. This is a problem at the district level.

Scott Marion No one assumes that a principal would have enough content knowledge. A suggestion is to have content area teams set objectives across teachers.

Laura Goe Yes this is what districts should be doing to have comparability in their schools. This solves the comparability issue on one level, but currently we only see this maybe at the school level and it should be done on a larger scale. This does take a lot of weight off of the state.

Paul Nichols What kind of quality do you see in these?

Laura Goe I have not seen them, but I imagine that it is rubric based, and this is allowed.

Paul Nichols But rubrics have a wide range of quality.

Laura Goe This is true, you will see a wide range of variability. You have to think about who shoulders the responsibility of creating this type of assessment, and what level of comparability they should have. No one is doing a perfect job of this. SLOs require pre and post assessments, but there is no understanding of setting an appropriate goal and selecting an appropriate pretest.

Scott Marion Yes, right now they are not doing appropriate pretests. They are using proficiency from other tests instead. I was thinking of Toby's flow chart, and my thinking has evolved overtime, but the way many view student objectives is like Churchill viewed democracy. Teachers don't care who is requiring the assessments; just that someone is requiring them. Once we get out of the Front Range who will calculate growth? There are shortcomings of districts trying to calculate growth.

Laura Goe We start out with a muddle. We don't know if we are doing this right. Overtime we want to have a resource bank and allow teachers to do the assessments and keep the good ones in the resource bank. I was trying to keep from rebuilding the wheel in New York. One teacher made a great assessment for sign language and she has no way of getting this out to others, but it should be used on a national scale. This will not happen overnight, but I really want to see a resource bank.

Todd Morse Regarding SB191, what if the only thing we accomplish is that teachers get better at assessment? This would be a huge victory.

Scott Marion We want teachers to set meaningful goals. I used to see SLOs as a last resort, but I don't anymore.

Toby King There is a chance for us to go to the board and say what we knew before and what we know now in order to keep the vision going in a meaningful way.

Kristen Huff Reiterated the case of New York: in New York we have over 700 districts, and BOCES district collaborative provides services. The human capital that goes into this is amazing. I want to caution Colorado against something we did; we need to get growth out of the conversation because it confuses people. As a Race to the Top state, we try to provide assistance to network teams across the state around SLOs and the way that we are talking about it now has improved. This is what you do anyway as leaders, you assess where your students are and you have a goal. It is different for every student. We have really tried to change our focus and give teachers more power. I want to underscore the human capital this all requires.

David Webb Let's return to the grain size issue on growth. Writing can be documented over the years. Are math, dance, etc going to be treated in the same way? How do performance and disciplines develop over time?

Scott Marion How do we craft these objectives overtime? I think many are trying to look at this and define it, but what about things that only last 3weeks or so? Time can be looked at in different ways too. Maybe we should see if we could try to write objectives, I tried this with others and it wasn't very successful. I hope by the end of summer I will see some great examples. What makes something an objective and not a bullet point?

David Webb Is a high quality assessment going to lead to a high quality measure?

Guillermo Solano-Flores I want to go back to the need to develop capacity. I think we need to think more about this and incorporate it. Social participation is interesting, I have found that even if teachers taught for 10-15 years, they have never developed an assessment (agreement in the room). My conclusion is that teachers don't believe that assessments are something that they can do, they don't see this as something that belongs to them, they see it as something that is created externally. We need to teach teachers that this is something they should be doing. As we think about how the bank is going to be created and used, we need to think about integrating teachers into taking an active role in assessments. We need to think about professional development for teachers.

Jacqueline S. Law Teachers don't understand data and how it is useful, so this is not surprising.

Toby King In Colorado, we have opportunities to take what we know and choose a direction. We have adopted the common course codes and if someone teaches a course maybe they click on the course and there are the SLOs. Then maybe a teacher could comment on the SLO or they could create his or her own and put it under the course. We need to teach teachers what these are intended to measure. We could then allow teachers to pick what works for them.

Jo O'Brien Think about the Harvard movie, Private Universe, if this whole thing does nothing more than allow educators to utilize assessments well, then maybe we do something where teachers can have a shared expectation, and we allow teachers to have their own beliefs, but we clarify what we mean by say Algebra I means this in all areas of Colorado. It would be nice to have a great picture of what we mean by an endgame for the outcome we are looking for once a student completes Algebra I.

Toby King We do have GLEs that can do some of these things.

Jo O'Brien Establishing a crosswalk, for say 3rd grade dance state-by-state standards, would allow us to double and triple down on a bank where we can eliminate some of the time investments and say this is what the common core looks like across the nation. There should be at a minimum a common agreement on these.

Laura Goe SLOs were initially about every teacher reinventing the wheel. Now I have a vision of a national bank. Race to the Top, could share results across a content, like middle school math, we could have a team of these educators get together and set shared objectives. This creates collaboration, mutual support, etc. Value added sometimes depends a lot on what the teacher taught the year before. If teachers work together then the content could build overtime. This may make sense in rural districts as well.

Sue Bechard Progress of the process has a lot of information that we can look at and apply towards kids. RtI already developed procedures: kids need to be identified and their progress followed over the year, I imagine Colorado has some idea of what this looks like, then you perform and see if the students meet their targets. You already have stuff that people have been using and this might make you a few steps ahead of the game.

David Webb Learning trajectories, learning progressions and the grain size issue is another thing for us to think about, and if we are really looking for growth, we need to keep the big picture in mind. Maybe growth is not the right metaphor, learning overtime has to really examine the grain size, I mean Algebra develops over the years. We need to ensure we do this with a little more clarity. If teachers engage in these conversations then they will have a better idea of how skills and concepts develop overtime. How do we measure growth? Some countries may have examples we could follow.

Sue Bechard Two G-SIG's, one involves 18 states, while the other one involves 13, and these are trying to develop a learning progression and common core map. Laying out the entire growth progression of K-12, where do you start and where is the ending point? The standards miss the layout of progression. Student learning objectives need to be clear to teachers and they cannot do this with out seeing the big picture of the long-term progression. Australia and New Zealand could be models to look at for these.

David Webb Why choose ambitious high quality assessments when we could focus on the more fleeting immediate method of recall. We could use accountability or incentives to get teachers to pick the more stringent method.

Timothy S. Brophy In Florida, we have a very different system then you do here, we work with standards and benchmarks. In our Race to the Top, process money was given to us to help measure hard to measure areas: dance, drama, music, etc… and we had 450 benchmarks once we cut them in half and weeded out a number of them that could not be assessed on a large scale. We have about 179 benchmarks we are working with this year with writer and reviewer codes, the DOE will eventually publish these and you should have access to these at that time. We have sample questions, and once they have been submitted, they go through many reviews. This structure should be available to you once it is published. NAFME-focused on reliability and validity and we struggled with the definitions of Race to the Top, we really struggled with what a highly effective teacher would be in these topics. We will be field testing over the next two years and we are trying to develop a growth model. The good news is that Florida's materials will be available to you all soon. This has a lot of buy-in from the field. Some teachers have biases but this is the nature of the game. We have 120 people working on this and it is nice to watch their development overtime. We are a big state so we will have some great data. I hope that Florida will eventually share some of our data with you.

Timothy S. Brophy We take advantage of every state conference to get the news out so you should also utilize these as networking opportunities to disseminate information to others.

Laura Goe We hire teachers to write our items, but this surprises teachers when you inform them of this fact. It is important to allow teachers to be a part of the overall process.

Reconvene 10:49am Design for field testing assessments in pilot districts

Angela Norlander, CDE and Elena Diaz-Bilello, Center for Assessment

Scott Marion This is a great discussion, I would love to talk more about student objectives, but I think we need to focus on the content collaboratives and the structuring of the resource bank.

Lets look at the first objective on page four of the agenda

Questions for TSC:

1. Should the technical quality evaluations of assessments included in the resource bank simply state that the assessment appears to have technical quality sufficient for measuring student achievement at a single point in time for a specific set of content standards?

Jo O'Brien To help you, I think Colorado has had an informal approach to looking at growth overtime. The word growth has been a sticking point and we may want to reframe this. Can you all give us advice as we decide on what goes into the bank?

Guillermo Solano-Flores I think what might be valuable, is to rate the effectiveness of an assessment based on whether it can be used in teaching. We want teachers to look at assessments as a way to help them reach their goals, but they need to do this through improving their teaching. We want teachers to have access to the tasks, but also to have information that will help them become better teachers.

Jo O'Brien One element of the assessments was, do they illuminate what is going on in the mind of the student? The pilot will also call the question of what do these assessments say about the methods and pedagogy of what is going on.

Guillermo Solano-Flores The evaluation forms are missing the element of whether the task is useful? Does the task make you learn something, or does it help you to see something in a new way? We want teachers to be critical users of these assessments.

Paul Nichols In order to use assessments in a formative way, we need teachers to take advantage of the information or they are useless. If we want to put a stamp on the assessments, then we should state that this has at this time met the standards of our review of the available assessment requirements and that this appears to have technical quality sufficient for measuring student achievement at a single point in time for a specific set of content standards.

Laura Goe I disagree.

Paul Nichols I think we need to state that our review of the evidence shows.

Laura Goe I think we need to steer clear with this wording, I don't like the use of technical quality in this suggestion.

Guillermo Solano-Flores Just because they show reliability and validity does not prevent them from misused. We need to have a teacher education plan that helps them develop and use this information.

Laura Goe We may need something like that later on, but not right now.

Paul Nichols I withdraw my previous statement.

Joyce Zurkowski I think we need to put some kind of approval on these with some degree of confidence. Validity is interpretation at the local level.

Toby King I hear the same thing, we need to state a certain level of approval, and after the pilots have positive outcomes then maybe we could put some kind of approval on them.

Jo O'Brien I don't know that a stamp is appropriate. These are items to be put into a body of evidence, where they are eligible for consideration, but we don't want one exam or one method. There are some of these assessments and measures that will be better than others.

Scott Marion I think stating the outcome and validity issues could be used, and then we avoid the use of technical language and stamps. Okay now the next step. *He referred to a document that was sent out to participants and asked what people thought of this document.* We have people that will be using these measures and we want to see what you think about these.

Todd Morse We have an idea that what we pilot next year will not meet the long-term requirements but that it will be a good starting point.

Jo O'Brien We do not have a formal campaign on assessment literacy at this time, but we are working on it. The summer symposium will introduce a shell of this and at the end of the summer we will know what this looks like.

Toby King I think we are ripe for introducing how to aggregate information on assessments and how they can be done. We have had 60 requests for training from the districts.

Laura Goe What guidance might the state provide for these? We have information about the resource bank, and we think they can measure growth at one point in time, but they could be used to measure growth overtime pretty easily. If you have a good rubric that is not task specific you may be able to use it to show growth. We need to identify the knowledge and skills that a kid should have and then look at the test for this at the end of the year. We could create different forms for this to test the metrics at earlier stages in the student's work. If the kid has a portfolio that shows growth overtime this might be more useful. We need the data first before we can show growth. Acceptable growth will be a value judgment.

Scott Marion Simple growth models do not work. I worry that pushing people towards growth will push people towards gain scores and these are not very good measures. I fear that people are going to do terrible things with these measures.

Laura Goe You have to get the data first and then see how it plays out before you start to figure out how acceptable growth is defined. You already have something that will measure non-tested and tested courses.

Scott Marion But we aren't really doing this, you can look in the paper I wrote. We need to shift now.

Paul Nichols But before we move on, we need to think about the system we are growing here. Let's think of the forest instead of the trees to think about how changes in one aspect of the system create room for another.

Sue Bechard One more comment, I worry about kids that are below and above our scales, and that we will not be able to capture their achievement.

Todd Morse Teachers will need to be able to measure these students.

Scott Marion Let's move on and talk about the pilot, collecting data on tasks and measures that go into the resource bank.

Toby King I don't know that the pilot is really limited. We didn't know the measures that would be found by the collaborative. The pilots want to try things that will not be held against them, and they want to keep an open dialog with us while they move through this process. Trish identified staff to coordinate and run the pilot. We will be working together with Trish and Jo on these groups. We are a little behind, but we are trying to move forward.

Scott Marion When we think of pilots we think of a large scale assessment with demographic samples, but we need to get that out of our head. We are working with volunteers, and we need a sampling framework, but are we getting feedback on how assessments work in certain places? Is it by item?

Elena Diaz-Bilello We are interested in collecting some data that has value.

Bill Bonk I think this goes back to the growth discussion, districts should be able to upload data and provide information that they can keep anonymous. We need to get some kind of quantitative data that we can compare and read. Is there a way to set boundaries with the pilots?

Toby King Yes, we need someone to design this protocol and we can support them.

Guillermo Solano-Flores In the case of reliability, what do we gain by saying this is sufficient? We need to distinguish what we mean by reliability. The technical property of the test may change as it is applied. The reliability obtained by an instrument may get us good data, but we may also have to state that using this assessment requires a high degree of training to get reliable data.

Todd Morse There were 80 boxes on the music grid, and if each of those has a test, that is a large assessment.

Guillermo Solano-Flores We will need a template with directions and judgments. Maybe a grant could be issued to find someone to do this step of the process?

Kristen Huff It is not feasible to get enough data from the pilots, so I would try to get as many assessments in the field as possible, and in lieu of doing focus groups you could do teacher teams with 2-5 people in order to provide feedback on their experiences. You could even do a survey. We might be trying to do too much with the pilots. Maybe the pilots should work to improve the rubrics.

Scott Marion What do we want to know from the pilots, Elena agrees that we cannot get high levels of quantitative psychometric data from the pilot.

Jo O'Brien We need advice that will help us avoid pitfalls. We know we cannot get high quality data from the assessments. How do we pilot an assessment, do we hand out seven of the same ones to see what happens with them, or do we pilot all of them thinly?

Toby King So we think we will have 7,000-9,000 teachers piloting these next year with a number of principals and schools. We need to organize and prioritize what we are looking for from these. The teachers are eager to contribute so far, and we can also get student responses.

Guillermo Solano-Flores An organization, the AAAS.org shows a number of questions that we may want to consider as we move forward, I will email this out to all of you. The NSF gave the AAAS a lot of money, and they developed these over a number of years. I think it could help us in our process.

David Webb There were some tasks under music that went across content areas, but it might also be nice to look across grade levels. The phase in is going to be important. It is difficult to improve rubrics without student work. So we need a wave of phase-ins before we go to scale across 9000 teachers in 27 districts. We are going to get bad results with the rubrics initially. So we should get this out to a few districts first, before moving on.

Jo O'Brien Teachers are also dependent on prior grade performance. So maybe looking at grade dependency would be a good starting point, or we could pick the pieces that need the most work on their rubrics and improve them before we go to the next round.

Scott Marion I like the idea of strengthening the rubrics first, and then looking across grade levels. This is an easier format than a psychometric approach. Then we are not adding content. We can only do this with open source stuff.

Jo O'Brien Is it a pilot or a construction task? If I am a part of a pilot, then I cannot select these on my own, I have to take what you give me, and then I have to improve the rubric. This is much different then what we set out to do initially.

Scott Marion but you also have a pilot in the second year, right?

Tricia Miller I remain unclear on what the goal of the pilot is at this point. We have a number of options to consider especially since we are supposed to start in the fall of 2012.

Scott Marion The Technical Steering Committee is the last group to tell you what the purpose of the pilot will be used to achieve. I think we are pretty clear that we do not have the horses for a psychometric approach. Tasks are good so far, but the rubrics need work, so I think this is really where we should start. We could say, you have to use this one, but you could also use a few others of your choice. This would help us to find a range of performance now with the pilots.

Toby King My teachers want to pilot assessments that will allow them to account for their 50% under SB191. If we give these to teachers, how can we allow teachers to use these to meet their 50%? How can we say yes or no?

Kristen Huff I think you can have two or three goals for the pilot. You could meet internally and prioritize goals and think about your samples. I think you can be flexible.

Scott Marion But we need to think about how these will interact with one another too.

Timothy S. Brophy I think it is good to go back to the stakes that are related to the results. If you are putting high stakes, like peoples jobs on the line, then you need to be more mindful.

Tricia Miller Yes, this makes me think of things that would be legally defensible.

Laura Goe So if you had to narrow the set of instruments, then you would need a strategy to figure out which ones are most important. Perhaps someone at the university could help determine this goal. It is pointless to pilot assessments that are not going to meet the requirements you are trying to meet.

Sue Bechard Can we collect feedback from teachers? We need to know how this looks when it is applied in the field. We need teacher and student engagement in this process.

David Webb I think some teachers would be willing to select anchor papers for the different measures.

Guillermo Solano-Flores SRI ten years ago or so had a project making assessments available to teachers.

Paul Nichols I think we want sample assessments with the least amount of evidence but the most amount of potential to show local and empirical evidence. You need to get student data, but you also need data from teachers and administrators. I think this can be done, but it will be a big task.

Sue Bechard We need to include all levels of students and different backgrounds. We probably want to have some prior level of achievement data on the kids before we use the assessments.

Tricia Miller This is set to go in September of 2012.

Toby King We can probably come together before then to assess where we are and discuss these issues a little further.

Jo O’Brien Thank you all for your feedback and your time. We will be in touch via email. We will meet here again in August.

Todd Morse Thank you all as I know this will help at least two districts.

12:01 Adjourn, Next meeting: Thursday, August 2nd