Nebraska Technical Advisory Committee Meeting
Nebraska Department of Education
January 31, 2022
8:30-12:30

8:30-8:45 Welcome and Introductions

Present: Chad Buckendahl, Jeff Nellhaus, Linda Poole, Cindy Gray, Christy Hovanetz,

Approval of Minutes from 5/27/21

Document 1: TAC Minutes 5-27-21 Approved by consent

8:45-9:45 Spring Analyses Plan for ELA and Mathematics

The Spring 2021 administration in the midst of the ongoing Covid impacts means the scale will need to be evaluated for stability by Spring 2022. NSEA recommend holding the cut scores constant for Spring 2022 while we review the stability of the NCSAS scale and the related linked RIT scores. To evaluate the stability of the NCSAS scale, NWEA will perform post-administration psychometric analyses by conducting horizontal equating for each grade.

Once pre- or post-equated solution for scoring is decided, lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) will be examined. From the 2021 NSCAS, we noticed that there were a larger than expected number of students who were received the LOSS+2 minimum score. Since the non-effortful response patterns are likely to result in the assigned LOSS+2 score, lowering the LOSS might be helpful in differentiating effortful but low-achieving students from those with non-effortful response patterns.

Final scores will be provided after evaluating the stability of the NCSAS scale for ELA and Mathematics. Before then, preliminary scores based on the current (i.e., pre-equated) item parameter estimates will be provided, with a note indicating that the preliminary score will be replaced with final score.

Document 2: NSCAS 2022 Spring Analyses Plan for ELA and Mathematics

- Does TAC have any other suggestions for post-equating checks for ELA and Mathematics? Jeremy gave summary of the experience and numbers from the Winter Pilot. Continued validation of results of the linking study. Chris Meador and Jungnam (psychometrics) presented plans for analysis. Shared 2022 test design.
- **TAC** – What do you mean by diagnostic? **NWEA**: Diagnostic items, could be off grade, will contribute only to the estimated RIT score not NSCAS accountability; decision made to make the accountability piece longer – want to give accountability determination enough weight; however, looking for ways to mitigate test length. Perhaps make Fall & Winter sessions shorter. Issue: passages in NSCAS longer than MAP G – NWEA/NDE look at what types of items causing timing concerns – considering lowering the number of items for fall & winter, but this will lessen the chance to give off grade items; **TAC**: Express concern NSCAS Growth will not reduce test time and will not get diagnostic information for interventions.
  NWEA: Post administration stability check will be through horizontal equating for each grade. **TAC**: What are you using for base year to make this horizontal equating? **NWEA**: We use the year when item was calibrated and for comparison will use pre- and post-equating. **TAC**: Did you have data pre-pandemic in bank & determine item parameters stability for the pandemic years? Do stability checks include stability of bank parameters?

NDE – Do not have operational data from last 2 years; stability check not available at least for the last two years. **TAC**: Idea scores reported for this spring will use post-equating results? Will you use pre- or post-equating? Will wait until post-equating checks are complete before release results? Yes. Will you do some or all before doing analysis? NWEA will wait until all finish testing. **TAC**: Are you planning to use winter pilot NSCAS Growth results for this? **NWEA**: Not exactly; winter will help with linking but not this pre-equating.

Scoring – provided after evaluating the stability of the NSCAS scale **TAC**: Release preliminary scores & then update them; if they are different is this a problem? May get a preliminary score then corrected later this only reduces faith in the instrument if scores are moving around; score for classification & RIT expectation comparable to MAP G – but schools are finding only 30% of the time will they fall within the estimated standard of measurement. **NDE** – After post-equating: expectation is minimal to no changes; with current system – actionable data w/in 72 hrs. running additional analysis allows the scores to be adjusted appropriately; communication is absolutely necessary regarding the process; believe numbers will be stable but need to do due diligence on quality checks. **TAC**: What score will parents get; will not be good if they get multiple scores. **NDE**: Summative/accountability determinations will only come after confirmation; **TAC**: Hold cut scores in place until NSCAS is fully stable? **NDE**: Yes – new cut scores will come for 22-23. Maintaining cut scores from 2017-18. **TAC**: If leave same cut scores with proficiency levels now, is message almost no student in the state is CCR? **NDE**: cut scores have not changed so expect 50% proficient as normal; will plan to change naming conventions for levels, but cut scores are determined by NE educators based on content not policy. AAAC weighed in on the content to use in that process. **TAC**: Understand the aspirational cut scores, but why NDE will wait to set them 2 years from now; LPS analysis – sending message 80-90% that they are not CCR. Districts concerned about practical implications of how high a student needs to score to even be considered proficient, let alone CCR Benchmark and political implications that the state needs vouchers to address student achievement. **NDE**: 2017-18 remain the same; political landscape changed. NDE not looking to change now because ELA standards would require new cut scores then do it again the next year for math. Expense is problematic. Intention is to keep things as stable as we can. **TAC** – Releasing two scores is unprecedented and confusion is problematic. Be very cautious with the messaging around two scores. Agree with maintaining 2017 expectations; need to be as accurate as possible. Ongoing issue has been high cut scores; having to score above 90% to become CCR. Need to rename highest level. **NDE**: Assessment will have national percentiles and confident in estimated RIT. We will do national norming with the state test. **TAC**: Is the strategy to link items to the bank and items then linked to the national norms. Are able to go back each time and continue to do checks on linking study and make sure the linking study is stable. Do parents receive percentile rank and performance scores? **NDE**: Yes, could have two percentiles. Reporting is always a challenge helping parents understand the numbers. Mindful of this. Will work with teachers on professional development regarding results. **TAC**: Do not reset cut scores earlier than planned. Hearing from other states that not letting other influences change policy. What does this mean for accountability in the state? What is the plan? Are you resetting the baseline, maintaining trend? **NDE**: We are figuring out how to maintain accountability system, so it is least disruptive. Since NDE will have a new Director of Accountability, NDE is trying to hold everything as stable as possible for established system. We have a working group with the Center of Assessment. They are forward thinking towards next year and new baseline scores. We would like to wait one more year, but federal and state lawmakers will not allow this. We will do the best we can with

the calculations we have. We are waiting for the federal government to put out a template. We will have many questions to bring to TAC. We expect to pull a theory of action and core tenets from the work with the Center for Assessment concerning an accountability system. We will pull back TAC to discuss accountability. **TAC**: Offer this – would not compromise on the test or expectations on students and make sure that you have strong testing as focus not the schools getting good marks in accountability. In Nebraska because accountability changed frequently that credibility of output are critical. Assessment stability is important not to mess with the measures. If can hold constant expectations in assessment system, then can adjust accountability as appropriate so the credibility is constant. **NDE**: Yes, keeping cut scores in place now is important. Reset with new standards/cut scores. **TAC**: One of perceptions is that assessment has not stayed constant. RIT scores change perception that assessment is changing. **NDE**: The basic blueprint and table of specifications has stayed constant. The basic blueprint is consistent and done checks for equating can show that consistent. Maybe have not done great job of communicating that. Have been getting question about accountability system. Educators are doing what they need to do. Will do the calculations with the data we have. **TAC**: What has changed is mode not substance or content.

Document 2: NSCAS 2022 Spring Analyses Plan for ELA and Mathematics

- Does TAC have any other suggestions for post-equating checks for ELA and Mathematics?
    a. **TAC**- this seems reasonable

- Does TAC have any suggestions for LOSS and HOSS adjustments for ELA and Mathematics?
  TAC suggestions for LOSS/HOSS adjustments? Issue & correction? Want to differentiate performance of the very low. Lower LOSS to account for effortful responses – differentiate from very low to just low. **NDE**: Many at bottom level but seemed to be great number of going through test quickly. If lower LOSS perhaps separate from low effort vs low performance. Will extend the scale so can differentiate better. Not a change in test design, except the student can go off grade level that does not continue to effect accountability score. **TAC**: Not sure how this will be done, the raw score would be the same for both students, so mechanically how to do that. **NDE**: Testing time can be used to help differentiate between the two; concern about room at the low end. Have several methods to differentiate. Also have practical concerns and is related to adaptive students because items are same for all students. Effort is valid. **TAC**: Given where median raw score is fairly below random guesses. Is not much space at lower end of the scale, do not have a problem adjusting the LOSS; but would like to see a comparison of using evaluation flag vs determining effort – response time approach/rapid guess. Then pulling them out to see if need to make adjustment. Given forensic checks (response time) that result in invalid score if lack of effort really leads to this score or uncertain score that cannot determine rather than assigning a score. Comingling of those who do not put forth the effort vs those trying to get a low score. Do we still have the same concern about the pileup of scores? Could say these students do not have a valid score and considered nonparticipants. **NDE**: Rather than adjusting scale have a policy change. We have not invalidated tests. Will allow retests for NSCAS Growth. Want valid scores and want to be more interim like in this way. **TAC**: Teachers do like ability to identify students who go through test too quickly and retest. Would like to see test progression. Would like to see something about the test determines retest, not teacher decision only. **NDE**: Guidelines in place and are tracking all retesting. **TAC**: Rather than change scale provide list of

students who appeared to click through. **NDE**: Have timing on back end, NDE has ability to track as well. **TAC**: Will need to do something mechanical to identify these students, so make it clear to schools which students have done this to inform actions.
NDE: TAC not generally against changing LOSS, but other approaches may be better. **NWEA**: Issue around median raw score - raises question, do we have easy enough items to accurately locate kids for spring. As look at fall & winter adjust test design to focus on diagnostic to get more precise RIT in tails instead of accountability and provide meaningful test results. **TAC**: Need sufficient on grade items to support peer review and accountability. Assessment needs to have capacity to support each cut score decisions. Put this up front. If density of scale is not enough in each location, then adjusting design (# of items in levels) in the locations is important. Conflating scale score adjustment & issue about motivation for students conflates two different questions. Do not have "heartburn" about adjustment, but do not know if adjustment answers the question. In high school, model from ACT to create consistent policies relating to score invalidation or retesting. Helps create consistent messaging. **NDE**: We have been slow because we do not have the evidence…been slow to react to limited flags to invalidate the test.
TAC: Considerations: Curious whether students can move through & go back through assessment? NDE: No, only go forward due to adaptability. **TAC**: Only 100 kids at LOSS+2; so be careful to make decision/adjustment to address this situation (few students); make sure not testing and retesting low performing students; if do institute this, how many schools/students are involved – relatively smallish problem – is this something we need to have a big definite solution now? **NDE**: For smaller schools could be big impact, but perhaps overreaction to scope of the problem.

9:45-10:00 Break

10:00-11:45 Plan for New NSCAS Science

The new NSCAS Science assessment has been designed to measure three-dimensional science learning, incorporating elements of Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs). Science was administered as a full-scale field test in Spring 2021. The dimensionality study confirmed that the unidimensional measurement model is sufficient to model Nebraska science assessment in order to monitor and report student learning progress in science[1]. Based on the fit statistics results, NWEA recommended the 1PL and PCM combination model approach, as this combination model not only fit the data well, but also provided more reasonable item difficulty parameters. Following the meeting between NDE and NWEA on September 22, 2021, the decision was made to move forward with the 1PL and PCM combination model and will reassess calibration model after the operational field test in 2022. Using the Spring 2022 data, NWEA will reassess the dimensionality and further investigate bi-factor model.

Once the measurement model is decided, Science items will be calibrated with the choice of model and student scores will be computed accordingly. If the decision is bifactor model, further investigation into scoring and reporting will be needed for score reporting in the following year. Then scaling, including scaling transformation constants, LOSS, and HOSS, will be discussed and determined.

[1] Nebraska Science 2021 Standalone Field Test Measurement Model 07-23-2021.pdf

The last step for new Science test is standard setting where cut scores are determined, which will occur in Summer 2022. Previously, NWEA used the Item-Descriptor (ID) Matching method, which was used for ELA cut score review and Mathematics standard setting in 2018. For science, the Range ALDs are more complex due to the multi-dimensional nature. In this case it may be more appropriate to focus on the bookmarking approach. NWEA would still use the opportunity to review the content in relation to the Range ALDs and make updates informed by data.

Document 3: Nebraska Science 2021 Standalone Field Test Measurement Model 07-23-2021.pdf

Document 4: NSCAS 2022 Science

1. Does TAC have any suggestions for science calibration? Does TAC recommend bi-factor model, given the item development based on three-dimensional science learning?

TAC: Are there only 18 questions and how many questions per domain? In grade 5, 18 points but Grade 8 has four 2-point items with 3-4 tasks. Surprised at providing summative score based on 18 points, when usually based on 40 points to classify into multiple performance levels. **NDE**: Trying to balance between testing time and item writing – tasks harder to write than standalone items normally see because of the three-dimensional nature of standards. **TAC**: Reconsider reporting at 3 levels. **NDE**: One suggestion we had was to only do one cut (proficient/not proficient) initially. With more data can move into more levels. **TAC**: How can we know they are curriculum content agnostic questions? If one student has access to topic & others to do not, provide them with real advantage. **NDE**: Phenomenon based and highly memorable. Based on Blueprint, all DCI (content Science) is part of standards and eligible every year and could filter through any phenomena. Students will engage in the content and be able to apply it. **TAC**: Can NDE check after the fact? **NDE**: Do not have a way to check OTL data, but we do have some information on curriculum used and instructional material used in state. We can do an analysis on whether this will play role. **TAC**: Regarding number of score points, “incredibly sparse” to make more than one decision point (mastery/not mastery). Looking at item maps, 5th grade range is very narrow and additional inferences very difficult (no capacity at upper or lower end to make additional decisions). Eighth grade has little more at top end, but nervous about making additional decisions. Characteristics of assessment will influence how TAC views psychometric model, standard setting practices and methodology. Please speak to longer term strategy. Is NDE intending to make decisions on these few score points? **NDE**: Expand number of items to choose from, especially at top and bottom of range, clear we need to go back. **TAC**: Need reliability with conditional SEM if you are going to have a scale, overall test standard error – on 18 points, difficult to see how get high reliability. Other thoughts: General agreement, multiple decision points will be difficult to achieve. How can identify grade level proficiency and how does this play in accountability system? If developing additional 100 items, are we building out assessment & how will affect expectations for proficiency and achievement? Continuing to change assessment. Will lose science growth data with this course of assessment? What are federal requirements around performance levels – at least 3 levels? If going 3 levels need more items on test. **NWEA**: Test worth time for teacher/student – this will be multiyear system around the blueprint and will speak to broader domain coverage. Have formative tasks available to teachers of high quality. **TAC**: You may need multiyear plan to fully implementing the program but as is cannot meet federal requirements. Cannot say because we could not get the quality of the test up high enough, fast enough you cannot compromise on reliability of scores. You are pushing for high validity but need high reliability as well. Knowing a bit about what state is doing, member is less concerned about substance

of science items. ELA & Math are adaptive and have more items, but science content standards are larger in complexity and scope. Seems counterintuitive you can measure science with fewer items. Plan should be a multiyear, but in any one year see representation given the breadth & depth of the domains of science. Schools should not play guessing game around which standards will show up on test. Sampling should make this possible.

2. Does TAC have further input on the method of standard setting?

TAC: Psychometric model - Have an underlying expectation of testing is have an unidimensional construct, but next generation science standards threw a wrench in with three dimensional language. Pelligrino said he did not mean it to be empirically multi-dimensional. Approach that NWEA is looking at is appropriate: Is the assessment ultimately multi-dimensional and needs that model to report scores. This approach taken has been seen in other programs. Is good starting point. Hypothesis that will be empirically being unidimensional but be able to respond. This would be a necessary source of evidence that can contradict claim that not addressing multidimensional. **NDE**: say is all one dimension in end because all science. Trying to make sure doing justice to standards. This is a tension created. **TAC**: Is there preliminary analysis showing that dimensions correlating highly with each other? **NDE**: Yes, analysis from field test indicates overwhelmingly unidimensionality works fine. **TAC**: Question is what measurement model will be chosen - one or two parameter models, Rasch, single dimension model? Is there a proposed measurement model based on preliminary analysis? From report used 1 PL with PCM for the one parameter partial credit model for the polytomous items on both assessments. **NWEA**: We compared 1 PL PCM vs 2 PL PCM and to VPN to PCN. Recommendation is 1 PL w/ PCM combination. Will do same FIT statistics comparison between the PL combination and one pair combination and investigate bifactor model.

TAC: Feedback is to go ahead & replicate dimensionality study with spring data & hope it shows same result to support confidence moving forward. [The design of the science test is shared by NDE &NWEA.] **TAC**: Wonders given the way tasks are constructed, if testlet analyses (item sets) may be more appropriate. This type of analysis is historically popular. Sometimes look at test sets information to look at dependency. Some items in set may violate an assumption of local independence. For instance, particular topic one district spent time on studying but another did not, the level of engagement for students on this topic is different. Result is set of items impacted not just one. Sometimes looking at this item set to determine if one item not functioning well, can replace an item within the set without losing the stimulus of this without impacting whole set. Build the testlet with more items so can lose one if necessary. Allows for attrition and/or replacement items. Extend shelf-life. **NDE**: easier for ELA than for science in this case. Some items in tasks based on what comes before – harder to generate more tasks to accomplish this. Risk of dependency makes case of more items. **TAC**: As explore models, you can review licensure and certification in CPA and some medical programs can be example of how to set up firewall to provide answer so they can move forward on test with correct information. Happens in real life.

On topic of scaling, NDE do we put all on same or use 3000 for science? **TAC**: ELA & Math have different scales so should science be different? **NDE**: each scale ends at different point as well. **TAC**: Concern is parents who try to use them interchangeably and draw conclusions that their child is better in math than ELA. Can you make scale, range, and cut score the same? **NDE**: cannot do this because we did a vertical scale so cut scores are different for each. We did this so we could show growth. Science is also different because we are only assessing two grades (5^{th} and 8^{th}). It is difficult to interpret with only a scale. **TAC**: Some

concern that parents will be confused and discount the test. Another TAC member stated if continue the system and remain on the vertical scale makes sense; their concern is making sure the 8th grade scores are higher than 5th grade so it makes some sense. Continue what parents are used to and provide context on student report is more important. TAC member asked would we recommend the NDE anchoring the cut score at the same level for each discipline? Yes, if possible. It will build an understanding by the general public on what score means proficient; however, difficult on a vertical scale. It is challenging, especially given the limited number of score points in science. Question asked of assessment advisory group was to increase the categories of levels to see more movement in scores. Cannot do this for science, but can we do this for ELA & Math? No answer. **TAC**: Can you hold off on standard setting in science for another year? **NDE**: Cannot hold off due to waiver given by feds for the past two years. **TAC**: With assumption NDE will go forward, would NDE be given the opportunity to make a proficient/not proficient in first year? Nervous about setting multiple cut scores on the number of score points NDE has. Have an interim cut then follow up with validation anchoring off the cut or cuts establish on interim (2 or 3 states taken this approach). Redesigning science (NGSS) when the pandemic hit. NDE will take back this information and develop a plan going forward.

NDE: Should we continue with Item-Descriptor Match (ID) or is it better to use another method? Because of ALDs and principle-design approach, we designed all tasks off these. Is Angoff or Bookmark better? **TAC**: In general, different methodologies will impact the recommendations due to different underlying assumptions, different types of feedback data presented, and different cognitive tasks. Do all 3 have impact data included at the end of the process? Built in some place. Misrepresentations for Modified Angoff method – all need target student before making judgments. So developed ALD then wrote items to elicit responses at each level. If this is case, they have preliminary idea of how student is performing. So standard setting is verification of these assumptions. Standard setting should confirm or not confirm of ALDs. **NWEA**: Will not have good coverage for each proficiency level for each ALD. We lost some through attrition. Item writers were told to write according to indicator, level (CCR Benchmark, Developing, On Track), and what ALD. **TAC**: NAEP took method like Angoff with instructions to implement different ways. Implementation can make a difference (not much), but distribution of student performance may result in large differences in proportion of proficiency. Impact data helps adjust cut scores when needed, but you want to make sure systematic process is followed. For these particular methodologies, they are human judgment based, but there are empirical based methods (like ACT). For these kinds of test-based methodologies, NDE is asking panel to intersect policy, content & empirical data to make informed judgments. What is typically recommended, is to use methodology that aligns with the type of assessment. For science MA is suggesting there is not an order effect, but there is (i.e., order in which the items are presented to the student). For Bookmark or ID, you are disentangling assessment & ordering score points consistent for easiest to more difficult. Intention is to reduce the cognitive complexity for panelists. For this assessment TAC is not fan of any of these methodologies. The testlet configuration makes others better: extended Angoff, dominant profile judgment (for complex assessments), or MAP Mark (NAEP). Each addresses uniqueness of test when clustering around phenomenon. Panelists think about testlet organization with multiple tasks/pieces. Not apposed to ID Matching, can provide some ease for panelists but may lose something in terms of how items are clustered together in terms of what will be expected for each cluster = becomes total and evaluate reasonableness of total. Bookmark is very common. It is important to have human judgment as part of the standard setting process, but implementation of the process and use of committees are key. **NWEA**: Given the limitations of the operational portions of the test forms, we should consider the methodology that can

leverage the field test items rather than just look at the distribution of the scores. Does this make sense? **TAC**: Your question is because of the potential lack of density of the score scale, methodology that would supplement with field test items that could build out greater density along the score scale for the panel to review. **NWEA**: Introduces risk of cut score outside the range of the test. **TAC:** Something to consider; if tasks and items to supplement scale for full density it may give you enough information to take this back to form construction and then build something that helps to maximize decision consistency around cut scores. Capacity of the assessment to support the cuts will be important. Without this, may only be able to support one cut given data we have seen. Right now could only support one cut on existing form length; not confident with low enough conditional standard error that you would be confident in classification decisions. Caution: looking at impact data – not normal instructional years. Nervous moving forward with including impact data in determination of cut scores. Relying more on educator judgment and what you were intending to assess and what you expect students to be able to do rather than what students actually did is more important this year than has been in the past. However, while most of Nebraska was business as usual, science did get more impacted than reading and math. Students would have had impact on opportunity to learn, even when school was in session. Trying to help people understand why states would return to cut scores and verifying them will be difficult. **TAC**: Aware of a couple of states, because of risk of overinterpreting pandemic era data, have established interim cut scores without impact data with commitment to review them when back to time of "normal instructional practices". Mitigation strategies to avoid pandemic slide perception: do not mitigate COVID slide and just state it. Students will not be as well prepared this year and really will not know for three years. Can make adjustments on accountability side. Needs to have some aspiration when setting cut scores. Concern that data will interplay with expectations. Other concern is aspirational cut scores may lead to students not taking science courses because had a lower-than-expected proficiency score. Communication is key here. NAEP prior to pandemic might be a good mitigation of baseline; if perception that previous science assessment was not rigorous enough, then would not use archival science assessment data for this. Would not fit well with step up in design of new science assessment.

11:45-12:00 Break

12:00-12:30 Plan for NSCAS Alternate Science Standard Setting

The new NSCAS Alternate Science assessment has been designed to measure extended College and Career Ready Science Standards for students with significant cognitive disabilities. In Summer 2022, a standard setting will be conducted for the Alternate Science assessments. Prior to the workshop, DRC will engage special education practitioners in a virtual ALD development activity. DRC recommends a committee of 24-36 educators be convened for the standard setting, with one group formed per tested grade. The Angoff Yes/No method is suggested as part of a two-day, in-person workshop.

Document 5: Nebraska Alt Science Standard setting Design 1-25-2022.pdf

1. What benchmarked test data, if any, may be appropriate for standard setting participants to consider as part of the standard setting?
   TAC: Curious about previous alternate science results. Would they have the same kind of challenges (they are so different not really want to use as benchmark)? **NDE**: Yes, there were large changes in standards. Even writing the extended indicators was difficult. Did good job of making attainable/achievable standards but is much more rigor in these standards. **TAC**: Would ELA and math be better points of benchmarking interpretation or

comparability for science alternate? Think of alternate as a system (part of ELA and Math results). Look at recent graduates from Nebraska high schools and see what they are doing and look at how they did on science assessments in the past. Can we make predictions about a score level, can you gain meaningful employment? That kind of research would be useful information. **NDE**: Would be great…not sure could do this. **TAC**: This would be evidence. **NDE**: Extended standards do a good job of real-world applications…not just theoretical but operationalize concepts for day-to-day lives of students. A lot of students are being served until they are 21. Could be contacted to "find out how they felt about things". **DRC**: It may not be something can do this year but will have to collect validity evidence for years and can do moving forward. **TAC**: Could include employers on standard setting committees for 11th grade. **NDE**: Specialized programs may be good – project search coaches. **TAC**: Will you use impact data? **DRC**: Will use electronic tools during standard setting. DRC will have three rounds of judgment with discussions of recommendations in groups then independent. Impact data will be shown after the second round. Gives options as to when to show benchmarks. Showing impact data will have challenges. Some teachers will be tempted based on pandemic. Use extended indicators as main guide. When show impact data will give advice, primarily around the indicators and content. Also plan to share what math or ELA scores would do relative to cut scores. **TAC**: On impact data, you mentioned concerns regarding stability of item parameters, do you have any concerns about sharing the impact data from a stability perspective (i.e., stability of the difficulty information)? **DRC**: Is a challenge of sharing impact data from such a small population of students. Due to COVID effect, way to frame impact data – we do not know if data will be generalizable for future years, but this is the best data we have right now. Possibly encourage participants to put less emphasis on this data – share impact data perhaps later in the process…more so with policy group. **TAC**: Benchmarks mentioned previously, will you potentially bring in data from ELA and math here? **DRC**: It would come in here to contextualize impact data. As mentioned previously, might push impact data to post-workshop policy group so the participants only work with content. **TAC**: NDE implied descriptors have changed, but some students on alternate assessment would be categorized as CCR. May have psychological impact on impact data if presented this way. Difficult to say a student in a life skills class is CCR, but students who do very well on nationally normed tests do not. **NDE**: Typically speaking, the proficiency levels have mirrored general assessments. This is generally an expectation from federal peer review. The two lines of assessment should share/mirror the levels. Do we want to policy-wise, but do acknowledge can be misinterpretations what this means for students? Need to have the right people in the room to understand the students who take this test. **TAC**: Naming of the achievement levels should stick to demonstrate the standards – did not meet, met, or exceeds. When you get into predictive information (like CCR, ready for the next level), unless you have done the research to make claim it is difficult. **NDE**: Understand to downplay the predictive information. **DRC**: Do plan a cut score review after the standard setting. Goal to review recommendations and, if needed, to recommend adjustments to scores. This is an explicit policy-based process designed to interpret the recommendations from the standard setting. Goal is to look at recommendations, talk about how similar/different than previously identified benchmarks, talk about why they exist, and make recommendation. **NDE**: Recommendation will go the State Board for final approval. **TAC**: Around this process there is a measurement error around process itself or with the test itself. The range of cut scores that policy group can make should be defined by the error in the process. This makes it more defensible. **NDE**: Important to be transparent about how cut scores are developed.