**Nebraska Technical Advisory Committee Meeting**

**Nebraska Department of Education**

**January 12, 2023**

**12:00 – 4:00 p.m.**

**12:00-12:10 Welcome and Introductions**

**12:10-2:00 Evidence of Validity for NSCAS Growth**

**2:00-2:15 Break**

**2:15-3:15 Evidence of Validity for NSCAS Growth**

**3:15-4:00 Standard Setting Plans NSCAS Alternate**

In spring 2023, NDE will sponsor a standard setting for Alternate ELA and math, a standards validation for Alternate science, and a policy review encompassing both. DRC has prepared a design document for TAC review that describes the planned four-part standard setting design. In addition, two questions are posed to the TAC:

* What is the best way for NDE to frame the 2023 math standard setting, given another standard setting will be needed in 2024?
* How should NDE prepare for work with the Assessment and Accountability Advisory Committee before and after the workshops?

Trudy K Clark, the Assistant Director of Statewide Assessment, provided a status report of NSCAS Growth. The test design for NSCAS Growth has changed from its original design (i.e., proficiency and growth score, 3 times per year). The assessment now has a more robust diagnostic portion for fall and winter. This means it is no longer guaranteed that students will test on the 27 items necessary to obtain a valid proficiency score in the fall and winter, 2022.

NWEA presents on the evidence of validity: interpretations of how scores could/should be used. Opportunity to how assessment is intended to be used and information specifically aligns with federal peer review guidelines for assessment around validity, reliability, & fairness. Peer review guidelines are based on the Standards for Educational and Psychological testing – industry expectations for assessment. Does NWEA have evidence that an assessment given in the fall or winter generates a valid proficiency score? In turn, USDOE will evaluate Nebraska’s assessment system and whether the evidence collectively support the claim.

TAC questions prior to presentation: Testing students in fall on standards that haven’t been taught yet – is it fair? Growth scores do not show what needed to guide instruction. Can proficiency claims be made before instruction met? Chance to reset what through-year assessment mean and how should scores be used from each administration. Acknowledging limitations regarding what it can/cannot be used for. Also there has been no claim for prediction for what students will do in the spring. What are uses NE wants districts to consider, what is appropriate, and what does the evidence actually support.

Clarification: Metrics/data generated during the school year assessment is different than what is generated at the end of year – score reports are different? NSCAS Growth does provide student reports intended on same scale and make claims for proficiency. One of the open-ended questions is the interpretability of fall & winter scores in comparison to the spring assessment.

Reporting piece is part of validity evidence, comparing & contrasting fall/winter vs spring administration.

NWEA presentation on 5 sources of validity. Clarifying present test design is not a through year since assessment is not banking proficiency. Validity is based on test results use.

Questions NWEA asks of TAC:

Additional validity evidence?

Which one of proposed test designs recommend?

What is Linked RIT score? Through linking we put the NSCAS RIT score on same scale as MAP Growth RIT. With the MAP can compare nationally. NSCAS Growth RIT cannot be compared nationally, correct? It is comparable.

Test Content: Did the test development process follow a rational approach that ensures that the content matches the test specifications and ensures that the test items represent appropriate knowledge, behaviors, and skills? TAC: States have external 3rd party beyond local educators conduct the alignment study, has NE contracted for alignment study? Yes. Talk about the number of questions you need or what you decided upon to show evidence that student met guidelines for test? For blueprint 26 operational items aligned to summative portion + 13 diagnostic items. One reason why NWEA is able to measure content standards with what may seem like relatively few items is where the adaptive nature of test comes into play. Be more efficient in targeting difficulty to student based on where student is in terms of their ability, therefore don’t have to ask more questions. Reliability is there. Size of item pool? Overall item pool for ELA & Math 15-16,000; 4000 of these are summative & rest is MAP Growth item pool.

Clear statement of Knowledge, Behaviors, & skills represented by test? Research? Is this different with full summative blueprint? Correct, only see full blueprint in Spring. Relaxed criteria to provide more diagnostic/growth information for remainder of this year. From discussions, concern continues to be fall & winter diagnostic questions are such a small number of items – tails, what schools use for interventions and for HAL students, not good. Are we trying to do too many things in one test fall/winter quite a bit to accomplish – should schools use MAP Growth so have evidence for interventions? Main reason for TAC input in test design – test design to prioritize – diagnostic or accountability. What was composition of expert panels? TAC suggested adding to participants list, their content area expertise for peer review, specifically to what degree are they content specialists (i.e., degree earned).

Test development process: TAC: Both fixed form & CAT are fed through the operational administration. Are you evaluating and scoring the fixed forms also using estimated CAT scoring or a separate, parallel process? Separate & parallel process for fixed form. Range ALD is not reflected in this process; you might consider including these in test development process. This is a way to present content standards and ALDS is a way to extend them. Since you use them in a more principled item development process, you want to add it in to strengthen process. You may also want to add a Content & Bias Review following Item Development to highlight the QC process that was described.

Bias/Sensitivity review: Do you do this review for ELA passages prior to developing questions for them? Yes, part of our passage stimulus review process. That is a separate meeting of bias & sensitivity committee? Yes, it is.

Item Tryout & Statistical Review TAC: Item difficulty – you get criticism if remove items too hard. Might raise question about the standard itself such as not well aligned. But if aligned with standard may consider rewriting item so it is more accessible. Be careful & not immediately reject an item because it’s too hard/too challenging. We want high standards. This is something to consider as part of your process & explain how it is handled. Inverse of that is noted for p-value below .2 or above .9. Same concept if too easy. We can blend this idea of how content intersects with data because content committee can look at RALDs in context of statistical review. We want the CAT to have sufficient representation of content regardless of ability so students can demonstrate where they are. Data serves as flag not final decision maker

Alignment: TAC: Does math & ELA go through standard setting for the general tests? Yes, correct. Both are planned for summer 2023. ELA to new standards, math to the legacy standards. Have any items & item content been created specifically for NSCAS? What item types are on the assessment & is there a major difference in validity between some of the item types? Should there be additional item development to be more in line with innovated item types? This goes back to item difficulty conversation…looking at more innovative types tend to be more difficult. Be sure not excluding them because of their statistics – look at item correlation (good proxy of validity). Step back and ask are we looking at this now because NE is submitting for peer review or is there a real urgent reason why? Part of the reason we are reviewing this especially in the framework of peer review is there are actually questions about design that TAC can provide input. There have been different perspectives on what scores are available and how they should be used. How NE could be defining what they mean by through year. Part of question is for NWEA to talk through evidence & provide suggestions on design based on evidence. If 3rd party completed an alignment and made suggestions on how to improve alignment, NDE will need to respond to suggestions. Clarifying question: With spring assessment, is there more grade level items…is the number of questions the same or different compared to the fall or winter? Constraints are tighter in spring & do field testing in spring – see a few more items, but constraints tighter so meet blueprint – not focusing on diagnostic items.

Response Processes: Test Engagement Level – TAC; MAP G pauses, but NSCAS does not pause, from teacher perspective good tool. Because only showing all students, issues may be masked for subgroups. In looking at peer review guidance, part of this evidence is whether there is alignment between method of measurement and content standards. Another important source is whether item types are appropriate given content standard. Include this in documentation and expect to see how is assessment eliciting intended response processes from students in representing construct. Distinguish between content and cognitive processes or cognitive complexity. Looking for alignment between standards, ALDs, and cognitive processes and are you using right item types. Make sure have complexity in item types we didn’t have previously. Within section want to bring in, including statistical sources & qualitative sources (item development and test design could fit here) Tests on computers bring up mode effect (i.e., would students score the same on a paper test vs CAT). Try to demonstrate mode didn’t impact student results but don’t have this here since NE students mostly on computer.

Internal structure: Support for subscore reporting. TAC: These are expected for a correlation matrix but curious given the strong positive correlations between reading & math, is this a flag that reading is important factor within math assessment and does this raise concerns about construct irrelevant variance. Without looking at factor analysis or dimensionality estimate, still anticipate reading would likely effectively unidimensional. When you correct for measurement error, does raise question since higher than typically see you are correcting for this. This is not a potential problem but claim of unidimensionality and see high correlations, it does raise questions.

Intended test dimensions – percentage of variance explained vs ratio of item/1st contrast. TAC: Suggests across content areas explained variance is similar, but science explains more but as compared to other, would larger numbers explain stronger unidimensionality? From an analogy standpoint, is this equivalent to ratio of eigenvalue? Yes.

Factor Analysis & Coefficient Alpha marginal reliability is reasonable values

Relations to Other Variables/assessments: Correlation between MAP G to NSCAS scores. This is the Spring NSCAS & Spring MAP Growth scores. To clarify Spring 2022 students who took both assessments during Spring. N = students who took both tests. There is a “fairly” high correlation between the 2 scores. These numbers are not disattenuated. NWEA Claim: While do not measure same construct, but really highly similar, we can treat them as identical and do a linking and confidence in comparability of RIT scores. TAC: Educators are ok with average students but concern is high & low ends. If use a spring test for placement of HAL or for interventions, can you do this from this test. Average makes sense, but concern is tails. If broke this chart out for high/low performers, would you see correlation numbers as high as we see now. Rather than full population, look at subgroups; you want to see impact for subgroups. Looking at number of items, so if on high end, wouldn’t correlate because they would not get same experience as MAP Growth. Highlights different design characteristics particularly item selection algorithm MAP vs NSCAS. You are trying to show NSCAS isn’t doing anything “funny” with measurement but consistent with other measures of student performance. You are arguing that the NSCAS and MAP are similar. Is there an opportunity to do a correlation with a different test? This is illustrating a student who scores high on NSCAS G will score high on MAP G & vice versa, but key nuance is constraints around summative measurement staying within grade level with proficiency but not a constraint for MAP G. From accountability perspective, users should have more confidence in NSCAS, but MAP Gr more confidence in placement. What is better test for placement? Use as much information as you can get. Look for consistent information across different measures might be the best way to make placement decisions.

Recognizing NSCAS component has to place ceiling & floor on to ensure most of measurement is occurring at grade level then need to factor this into interpretation and adding MAP G. What is purpose of each assessment? What do you want to get and how will it be used then go to design questions.

Consequences of Testing: TAC: MAP G is a diagnostic test of 40 items but when determining placement based on 16 items isn’t MAP G more accurate? NWEA RIT is based on all 40 items. Question continues to be if student is struggling and out of grade level, how can make decisions based on NSCAS G? NWEA agrees test cannot do both well. TAC: If making placement in Spring based on data in spring, which is better to use? This data shows not much difference. Maybe you don’t have to do both in the Spring. Maybe during the school year this is different, but if you are making decisions in the Spring administering both tests is unnecessary given strong similarities of results. Admin TAC: Tales is the issue because NSCAS G doesn’t address these students. Suggests guidance for field on how to use different tests; need a different meeting to have time to tease out this guidance for school districts.

Test Design: TAC: Other states using fall & winter diagnostic information as the starting value for the adaptive algorithm for the spring summative assessment to try to reduce the test length for students. The system is using the fall/winter information as an a priori estimate of what part of the distribution may start the test…they won’t have to answer more questions than they need to get to the summative decision in the spring. Design you want is based on utility or expected use of each administration. Which design best serves the expected use. Create statement of purpose for each administration then can better understand what to suggest. Teacher TAC: Go for MAP G fall/winter then summative. BUROS did some work in alignment of MAPG to NE content standards. Is this still the case? If so, design 2 & 3 are effectively the same. Wouldn’t this then be a through-year assessment? Students get RIT in Spring correct? Could see improvement right from fall to spring? Right? Yes. In terms of information you are getting, will there be much push back from field if fall/winter only diagnostic? Teacher/Admin TAC – no push back, want design 3. TAC member - Looking at Design 2 & 3, concerns go back to item types and content constructs and how they are linked. If we have really good summative items, and not using rigorous items like you would see on fall/winter, are you getting the information you need to be ready for Spring. Are we testing the right type of things? Would like to see system with rigor, content and depth of what see in Spring throughout the year. Exposure to other item types so it is not a novelty when they experience them during Spring assessment. Will pick this up at future meeting, coalescing around design #2 & 3. What we are hearing is field is more comfortable with fall/winter MAP G instructionally supportive recognizing spring assessment is summative.

Alternate Standard Setting – Presentation by DRC (Data Recognition Corporation)

What is best way for NDE to frame the 2023 math standard setting, given another standard setting will be needed in 2024? TAC: this raises a red flag. If content standards have not changed, and ALDs have not changed, is policy rationale related to the pandemic? If so, just say we are lowering our standard. Problematic to say the content standards haven’t changed, but we don’t like the outcomes. This is how it is portrayed…that is what it sounds like. If field feels too rigorous, State Board should ask for more evidence. What are the other measures telling schools this? Commissioner/Board of Education policy decision and should be given evidence besides feelings. For other standard setting activities, unless wholesale change, may want to consider validation. Do not have enough information here to say recommended action is correct. Perhaps look at alignment of items to be sure test is aligned properly with standards. A review committee may be sufficient to review the ALD to determine if there has been enough change to warrant the new standard setting. With this population of special education, it is not easy to find benchmarks. Possibly use general education math test or number of students in the state scoring as proficient on the NAEP.

In discussing the standard setting process, TAC requested clarification on use of benchmark vs impact data. If you set the cut score this is the impact, or percentage of students that fall within the achievement levels. Benchmark: is this a reasonable percentage of proficient students based on an external criterion.

With respect to the math standard setting plan, this not a good idea. It will be looked at as a lowering of standards especially if it is in reaction to COVID. This action sets precedents for other standard settings. It calls into question judgment of educators, the process, and of the Board. Redoing these because of COVID is challenge. If already committed to standard setting, it should be a review and a bring back same exact educators that made the decisions and looked at through the lens, “did expectations change or performance change.” In this case, it appears to be a performance change. If you are going to do standard setting for both ELA and math, same information is going to be used to inform those and we do not want to see a lowering of expectations across the board. I see challenges in other states as well. I see knee jerk reactions to make COVID look not so bad and to get state performance back to what it used to be rather than navigating through the tough instructional work we have to do to get results back to where they were. We need to show policymakers that everything is not okay. We do not want them to stop dedicating resources and time and effort to policies. As far as working with AAAC, communicating with them prior to meeting to give them insight so they are not making knee jerk reactions. Do everything you can do to maintain 2023 standards. Be sure to include the 2018-19 data, make this as showcase rather than 2021-22 data.

How should NDE prepare for work with the AAAC before & after workshops? For benchmarks and then post review, there are two (2) large testing consortia for this student population, and might provide good benchmark data & help inform if the cut scores are too high. Use these as a comparison of NSCAS ALT to determine if the test is too stringent. May be worth bringing back original group to review the process to identify changes. For the AAAC, be sure representation across state for the review committee. When talk about benchmarks what other states utilize this? Pretty common for states to use NAEP for 4, 8, & HS for ELA and math. Also use MSAA and DLM for indicators of ALT assessments (assessments referenced earlier). These are external assessments. Policy committees may use a Hofstee and Beuk methodology to set a policy tolerance. If the work needs to happen, then find a valid reason. When going to AAAC, help them frame discussion by bringing criteria in as to what are professional standards & implied guidance (Critical elements) around standard setting. If you are bringing the cut scores to a policy group, question is how much they can change the cut scores determined by the teachers/process. Be sure it is data driven. If do not bring in 2018 teacher group, you could use the standard error of measurement from this work and take it to the policy makers. We want to accurately setting expectations for students with this population based on ALD and accurate classifications that serve these students,