Minutes – Nebraska Technical Advisory Committee meeting from March 22, 2020

**Location and Time:**

Virtual Meeting

March 13, 8:00-3:30

**Attendance**

**TAC Committee Present:** Chad Buckendahl – Chair, Cindy Gray, Robert Henson, Jeff Nellhaus, Linda Poole

**Approve Minutes September 20, 2019** – *Approved minutes as submitted*

**Introduction to Through-Year Adaptive Model**

* Jeremy Heneger provided an overview of the Through-Year model.
* All three (3) administrations give districts same information they get from MAP Growth plus ALDs (i.e., RIT scores, Learning Continuum & grade level information) – basically 3 interim tests
* Will use totality of testing information to create proficiency score at the end of the school year
  + Student experiences on test administrations 1 & 2 will dictate what the 3rd administration looks like –all test administrations will be adaptive
  + Schools and NDE want to get growth information for students
* NWEA research centered
  + Place items on the NSCAS scale and project to RIT scale so district get projected RIT & get placement on NSCAS scale
    - Attributes of item bank keeping comparability for summative piece
    - Item banks deep enough to measure growth on grade level primarily
    - When necessary & item bank too difficult, what is best method to go off grade & give instructional information to teachers? Trying to control when go off grade level in purposeful way.
  + Need to have consistent Blueprint for 3 administrations of test
* The goal is to provide a linked RIT score on NSCAS scale and keep NSCAS Blueprint
* Message to schools – teachers & administrators play role in how this TYA is portrayed to students – help in characterizing it as 3 interim tests – get MAP Growth results three times - not high stakes

***TAC***

* In terms of the roles of each test session, what role does each play and what metrics are reported?
* Will students get a performance level & scale score?
* Will 1st & 2nd tests serve as locator tests? In terms of the adaptive nature, the student begins the next administration from the point he/she ended the previous test.
* Are 3 test administrations equivalent or will 3rd become summative with proficiency only based on Spring test? What is the summative information reported in the first 2 administrations-desirable options/metrics?
  + NDE – proficiency based on whole of it; NWEA –– want to integrate two testing systems & interpretations of what students know & can do are on same scale – want to build this as a formative assessment construct: used to measure growth patterns, but also identify achievement levels – looking at the validity evidence for the range ALDs – give accurate score interpretations & give help in formative framework
* If calibrating MAP Growth on NSCAS scale, is it not actually putting NSCAS in all administrations? Calibrating Growth items & modifying the Blueprint (i.e. changing Blueprint content)?
  + Keep NSCAS construct. Align MAP to NSCAS standards and range ALDs. Items are edited for style guide of NSCAS & item specifications to recalibrate content to NSCAS. Item bank is the point of research – one bank that binds all together or use three “mini-banks” we control over time.
  + Consider increasing complexity of MAP Growth items in order to measure accurately (due to growth on MAP scale but not NSCAS scale)
* See NSCAS Innovation design as huge breakthrough if works
* What sort of claims in messaging to schools with respect to how fall, winter, and spring should be used if summative aspects in each test? Encourage expectations that schools use this three times a year – how much choice is there? – aggregate will influence how schools use (more summative information you glean from earlier parts of the year, the more confounds exist in data – is it baseline student experience or instruction – do these account for the unevenness of scores)
  + Claims from each piece separately & aggregate of those claims are 4 separate validity arguments influence what you are expecting of NE educators, how they use the information in the classroom, & then how the accountability team uses the information.
    - From NWEA: Scores represent learning represented by the Blueprint & are calibrated in the Spring as end-of-year. Scores in fall & winter are intended to be more formative in interpretation even though they too are on Blueprint. Used for instruction rather than summative information.
    - Concern: spring administration needs to satisfy summative peer review requirements – constraint engine will have rules in place to satisfy these
    - Important element: students start new assessment where they left off from previous administration. Experience less material too hard or too easy, less testing time
* Intent to use fall and winter to get stronger formative information is better message. If they have summative components, could be more efficient/effective method to get summative results. Otherwise perception will be 3 state tests 3 times a year. Do not see discussion of creating profile. Compensatory vs conjunctive.
* Avoid fact taking 3 tests, most of language for fall & winter around expectations of where student will be in spring. Use language emphasizing formative nature on way to proficiency in spring. How much does proficiency in fall & winter go into determining proficiency in spring? Need to address this.
* Is the intent to bring MAP Growth onto NSCAS scale or creating an amalgamation or bringing NSCAS construct into MAP Growth scale through content and linking, since the secondary goal would be having within school district information but also connect to broader norm group of MAP Growth national participation.
  + NWEA: Relationship between MAP Growth and TY – TY design based on systemic perspective concerning how do you provide an assessment system in a way to encourage student learning (MAP Growth) but at same time satisfy requirements of summative test at EOY (NSCAS). Design goals are to allow students to test off-grade to allow teachers to meet them where they are instructionally and to provide data on growth across the years and within year. Working with NSCAS construct means doing linking studies to project the RIT and national norm.

.

**Linking Study between NSCAS and MAP Growth (Common Person Linking)**

* This linking study used IRT concurrent calibration to link NSCAS & MAP Growth scales based on the Spring 2019 data. Data sets were merged to create the sample.
  + What is meant by “same scale”? The mean b transformation constant is used to bring MAP Growth onto NSCAS scale
* Study designed to investigate degree to which MAP Growth items could be brought onto the NSCAS scale & achieve comparable results to NSCAS and be able to project MAP Growth RIT Score from NSCAS items.
* Three sets of scores produced/analyzed: projected MAP Growth scores from NSCAS scores using the mean b transformation constant; projected NSCAS scores from MAP Growth scores using the mean b transformation constant; MAPNE results obtained from separate scoring runs using MAP Growth items only while fixing the parameter estimates from MAPNE.
* Results:
  + High correlations between MAP Growth & MAPNE, NSCAS & MAPNE, RD & RDLU, and LU & RDLU.
  + Projected MAP Growth Math SD increases as grade level increases

***TAC****:*

* Pattern of differences between projected and actual scores more pronounced as increase in grade – is there anything in the data to suggest why
* MAP Growth goes off grade & NSCAS does not – difference in adaptivity
* Disparity in achievement levels as students age may be less motivation
* ***TAC***: Is there greater alignment between content and items as earlier grades?
* MAP Growth has different construct than NSCAS (different styles and permissions)
* Selection of content for MAP Growth may play part
  + Why is there such a big mean difference when use mean b transformation? All placed on same scale so measuring similar or same construct
* Standard deviations are very wide when you make the conversion unlike the item correlations
* MAP Growth & NSCAS use different algorithms in estimating Theta during administration. Math data has very long tails typically not seen on state assessments.
  + There is an equivalent group so strange that linking would result in estimated scores are so different.
* Different testing conditions for MAP Growth vs NSCAS
* Considering placing MAP Growth items in the NSCAS administration to see if it ameliorates what see in the data.
  + When moving to the RIT Scale is that a simple transformation of Theta?
    - Yes, taking mean difference.
* Through-year Common Person Linking Study Guiding Questions discussion – ***TAC*** *comments*
  + Hold off using equipercentile linking because there are too many unknowns
    - Secondary question asked: tell us more about how equipercentile linking would be used here
      * Ability to build concordance tables
      * Cuts are set where student has 50% chance of being proficient based on the state assessment
  + Correlations between MAP Growth and NSCAS are .8 -.85 (ELA .5, LU .8, Math .87)
  + If difference in how accurately measured & link them for may not expect same error
    - Marginal Reliability for MAP Growth is .95 but seeing in NSCAS .92 (Math) and .898 (ELA)
  + How does Blueprint compare to MAP Growth? May be big difference because measures different things – constructs different
    - Until MAP Growth more closely reflects NSCAS Blueprint will have problem with common person linking or common person
  + Because there is a discrepancy between the projected and actual scores, need more convergence before NDE uses message of MAP Growth/RIT projections.
    - Use 5% as margin of error if want confidence in projections (soft goal)
    - Need to understand why the numbers are as they are before using data to make projections.
    - If use is student intervention, err on side that error should be small
* MAP Growth Item Selection, Simulations, and (Common Item) Linking Plan

NWEA will conduct study based on MAP Growth items that are the least different in formatting from NSCAS and placed at the end of the Spring 2020 NSCAS test forms. Items are spiraled from pool and make the test slightly longer. Cognitive confusion over formatting differences would not affect operational scores as they are presented after the traditional NSCAS items. These items are not included for calculating student score. Similar procedure to common person linking.

***TAC****:*

* What is simulated? Demographics & # of items at the end of the test met constraints with 1000 students
* If items are similar in formatting, are they also similar in content? If they are, do they exclude the portion of MAP that caused the issues noted in the previous study?
* Same constraints for NSCAS now will likely be placed on NSCAS TY – business rules will be the same & field tested
* Selection process seems reasonable
* Cannot approve results when haven’t happened – NWEA will share simulation report as part of deploying engine (if NDE approves)
* Need to mitigate motivation effect
* Will students know questions will not count? (No)
* Will they be motivated?
* Effect on items appearing where they are may mitigate scores
* Monitor engagement / response time as students complete final MAP Growth item
* Adaptive nature of test
* If items really differ, how conditioned students are to experimental items needs consideration
* Style guide was more of an issue – cognitive shift was not a concern
* Population similar – items getting responses (1000-1500) would have been better
* Are there concerns about bias regarding student ability levels in sample?
* Important for representation across ability range
* In previous study data based on adaptive test- means any given item has truncated range of theta
* Could this explain results saw whenever there was a calibration

**Through-Year Research Studies for Year 4 (2020-2021)**

* NWEA - \* ongoing
  + Score comparability of MAP Growth on the CBE (Project Altair)\*
  + Linking study between NSCAS and MAP Growth\*
  + Simulations – goals: 1) Evaluate NSCAS item bank overage of unused items in operation to determine if sufficient numbers of items can be moved to a winter modified MAP Growth bank comprising NSCAS & MAP Growth items and achieve comparable scores to spring & comparable score interpretations to NSCAS Range ALDS; 2) Evaluate intended configurations for modified MAP Growth of going on- and off-grade.
    - Modified Growth bank will be created using NSCAS items & MAP Growth items aligned to NE content standards
    - Need to be able to give projected RIT score & NSCAS score and measure growth on 2 different scales
    - Estimate need 1500 items per grade for TY – augmenting banks with items from edited MAP Growth and field test in spring plus new item development
  + Modified MAP Growth Pilot – Winter 2020-21 Participating school students will take modified MAP Growth test that consists of NSCAS & MAP Growth items. Goal: pilot the functionality of the on-grade and off-grade business rules & the similarity of the TY results compared to both MAP Growth & NSCAS results from a trend score perspective
    - Need 2000-2500 student per grade level
    - Anticipated outcomes:
      * Mean scores should be similar to traditional MAP Growth trends
      * Average NSCAS mean scores higher than the lower adjacent grade spring scores & lower than same grade spring scores of previous year
  + Stability of TY to RIT Scores (Spring 2021)– All students who participated in spring NSCAS administration. Wil review trends fall RIT scores, winter, & spring with concorded winter & spring RIT after merging at student level; review trends for winter & spring NSCAS; investigate historic trend data from fall & winter from previous 2 years of data (2018-19, 2019-20) for participating schools & districts. Need to determine when to roll out the business rules for TY. Goal: Evaluate linked scores from TY to traditional MAP Growth

***TAC***

* Expect goal to have one large bank to provide support for all three (3) administrations or three (3) banks are linked
* Difficult to maintain multiple banks with links- single bank is easier to manage but needs to be large enough to meet any content specifications & constraints for exposure
* Stability of Through-Year Assessment to RIT Scores (Study 5)
* Contingencies if things do not work?
* Will have linking study and spring results to share in September, as well as comment on specifications
* Will have data to share in Feb following winter pilot
* Good studies
* Need quality data pool (High Quality) to pull this off – more information on how to build this especially with adaptive design
* Did not see literature review addressing longitudinal assessment within recertification (was not based in education field)
* Related to how to combine results to make conclusions/decisions – decision theory
* Some research in this space to mention (American Board of Internal Medicine & Pediatrics)
* Might help NE in TY
* AM Board of Anesthesia and AM Board of Internal Medicine published
* Look at these to draw methods from to add to any validity argument NWEA makes

**Spring 2020 NSCAS Science Field Test and Analysis Plan – (Update from pilot)**

* NWEA: A full-scale standalone field test Spring 2020 to determine how well the new science test design and performance tasks function prior to the operational launch in Spring 2021.
* Science Field Test Design measures three-dimensional science learning from the NCCRS-S. There are common tasks on three (3) field test forms (not linking)
  + Will operationalize one (1) test of comparable length with break and comparable coverage
  + Survey questions at the end of each test to standardize length and garner feedback from students.
  + Check on success of creating the randomly equal groups & context effects if have common tasks
* Data Analysis plan
  + Data integrity check – run for TRF & IRF data
  + Engine evaluation - CBE to verify all tasks, prompts & survey administration as designed – verify correct order
  + Check scores for tech-enhanced prompts
  + State representativeness of field test student sample
  + Response-time analyses – start & end times & time spent on each prompt
    - How much time to take the test – is it reasonable? How much time is unreasonable?
  + Prompt- and test-analyses
  + Survey question analyses
  + Continuation of the dimensionality study - Measurement Model Evaluation

***TAC***:

* Are you doing study to determine time for timed test? No
* Account for outliers – participants go too fast or too slow – on item basis or respect to average time? Average time per task
* Look at pilot test at individual item level for outliers especially if administered over multiple sessions – will have extremes on upper end if not closed out properly
* If use pilot test form longer than typical form response at item level may give information on form construction after the fact to balance time if cluster of items deviates from average expectation.
* Are students motivated enough? Will the students guess to get done? Are outliers identified as those who guess all the time?
* Look at test construction and tasks (long or short) to determine if designed as we want it
* Curious how NWEA is defining outliers – distribution is basis for this
* If student spends long time on test, do not include them?
* How much time is too long for relatively high-performing students do on test?
* Look at performance vs time
* Recommendations:
  + Some analyses prior to modeling analyses consistent with very specific beliefs – make the most sense if using total score and with multidimensional model
  + Caution – like descriptive statistics, how useful these are in interpretation depends on actual structure of data
  + Cognitive diagnostic model and continuous model – encourage NWEA to think more deliberately on how to approach things regarding belief what is out there (i.e., model comparison leads to belief)
    - What is first belief based on construct measuring and how used and then what types of comparisons will you use to confirm, support beliefs?
  + Very critical on structure and use of data and dimensionality is debated – use intended to provide scores for three (3) dimensions? This may influence model
  + Sequencing – model then analysis based on this model
* Results considered preliminary with post-equating/calibration?
  + Single reporting category – consider releasing some questions with item analysis results in first year at the school or student level
  + Help people understand the nature of standards and the expectations around the test
  + Consider validation investigation after 1st administration
* ***TAC*** guidance to field for use of similar science test from another state
  + Release item level data – be strategic – important in first years of new test
    - Fears about teacher using data inappropriately
    - Item analysis down to student level on every question & comparative data
    - Need training so don’t interpret incorrectly
    - Want test to inform instruction and analyze data (led to concrete conversations)
* ***TAC*** members would like to see item type samplers – NWEA will send to TAC members

**AQuESTT Accountability Topics**

**Review ESSA Designation Filter**

* Approximately one-third of NE schools qualify for designation of TSI/ATSI.

***TAC****:*  What is general sense of credibility for TSI/ATSI for last year?

* Feedback ranges but Lane took school administrators through the data so they saw how they ended up where they were. Less about credibility but rather not having the data in front of them.

***TAC***: Concerns include system is so convoluted that creates learned helplessness – so complicated administrators cannot understand it & no control over it.

* NDE states can do a better job of creating sense-making materials & having conversations with Lane
* The system is intentionally complex: tension between simplicity of process and fairness & so many indicators
* What makes this complicated is original intent/design around not ranking schools, but need to get classification levels – do not want non-educators using this data to rank schools
* Need to do better with presentation of data, timeliness, & how viewed by schools
* Next steps & what must TAC see from accountability to make changes/decisions
  + High schools under identified
    - Designation filter compared student groups in school to subsets of Title I schools (27) in the bottom quartile – Small comparison group
      * ***TAC***: as a reference point not broad enough - could use other SES indicators in combination.
      * Are there other indicators of success could be using?
    - Using graduation rate as an indicator – pretty high rate for our subsets
      * Used average - based on guidance from DOE – maybe use median
      * Once look at this figure, approach DOE to see if guidance changes

***TAC****:* If there are things missing, use qualitative identifiers for schools

* + - Find ways that schools would be included
    - Look for additional spaces of SES or other indicators for comparison in high school
    - Concerns regarding the verification of ratings/self-assessment & effect of EBA on designations
    - Use a two-year rolling average of ACT scores if schools do not have 10 students - increases pool of schools/students (schools must have at least 10 students who take the test)
* Filter methods for TSI/ATSI
  + Are our measures of growth adequately capturing progress?
    - Improvement, growth & non-proficiency
* ***TAC***: Are they making enough growth? What is enough growth? What is reasonable? Is there a minimum amount to demonstrate progress?
* Is Stage 2 irrelevant?
  + Not great amount of change for the 148 schools so is something wrong or is 148 schools at TSI/ATSI too large? Which situation is most concerning?
  + Concern for NDE
    - 276 schools not performing at the same level of Title I students but didn’t grow at the same rate
      * Need to know if true – if focus on growth, this is not happening in schools
    - ***TAC***: Is status highly correlated with academic progress/proficiency? Academic progress and non-proficiency is basically measuring the same thing – concern for NDE
      * If not seeing high correlation, may be less that progress measure colinear but the threshold expectation is issue
      * What does distribution of progress indicators look like – narrow or wide distribution need to shift threshold expectation to filter from Stage 2 or Stage 3 – change the cut score depending on shape & density of where the distribution is
      * If 148 schools at final stage too many, filter thresholds need to be adjusted if not enough confidence in results
* Comprehensive schools – concern about capacity & ability to support these schools especially for students with disabilities subset
  + Concern also about exit strategies for ATSI student group/schools
  + ***TAC***: Credibility issue if support does not result in improvement and schools may feel public shaming with no purpose if there’s no improvement
  + What is policy tolerance for State regarding identification of schools? Policy determines what is measured by accountability. Reality of intervention is whether you can afford to do it, but policy is what are you willing to say publicly. What are implications if you identify schools but cannot help them improve, and if you don’t identify them, but they need help? Consequences for over- and under-identifying schools. Get input from Commissioner and/or Board, if they are the ones who have a say in the process, regarding policy tolerance in order to get to point where you can push/pull on levers in system.
* Are Title I schools appropriate comparison group (lowest quarter for TSI & lowest 5% for ATSI)?
  + ***TAC***: Not sure if there is another category to use.
  + Is NDE required to stay with comparison group or can NDE change model (i.e., absolute threshold)?
    - Concern for NDE is when use absolutes it becomes easier for others to rank schools/students. Not willing to go there.
  + NDE: Goal is to identify schools but keep it extremely categorical & focus on those
    - ***TAC***: After years of data & model unchanged, schools may be able to control “own destiny” & get out of designation but if normative every year dependent on performance of subgroup difficult to gauge whether will be designated or not – issue with moving target if designated one year but not the next – look at criterion-based approach
    - System – look at references over past years & set it – if need to change it can come after re-evaluation (i.e., need to identify more schools) – become systematic so schools can identify when can drop out
    - Can look at schools designated over years & see what other schools you may have identified. Run logistics model to predict schools using compensatory equation. Problem would be if equation gets out schools will know what will keep them out of designation.

**Student Growth Percentile**

* Three measures of academic progress: improvement, non-proficiency, growth. Board is interested in emphasizing growth. Question: Should Student Growth Percentile be added to the accountability system?

***TAC***

* The simple gain score model has potential to fluctuate more based on sample characteristics than Student Growth Percentile (SGP).
* Other approaches should be used before SGP?
* Concern SGP not scientific or obtain good data (e.g. standard error measure)
* Multiple growth measures may be confusing – simulate accountability model with SGP rather than gain score – see if you get the same results in identifying schools for assistance.
* MA used SGP in accountability model & with training, educators can understand how SGP is used.
* If have SGP can create median growth percentile for school & scale distribution
* Cautions around SGP:
  + If not relative to target of meeting standards, people say you are giving credit for growth but not making growth mark – counterpoint – status already in model so acknowledging change rather than meeting a particular bar (messaging)
  + If have multiple indicators, difficult to explain & simulating impact on accountability model is important but also evaluation appropriate growth indicator with end to determine which one to use.
* Measures using trend SGP vertical table with values over longer time at school level – programmatic information rather than student level
* States shifted to SGP & how represent it – median is how many are doing it
* Some states use values table in part dependent on who helped build the model – advanced by Rich Hill