Nebraska Technical Advisory Committee Meeting
Nebraska Department of Education
May 27, 2021

TAC members in Attendance: Chad Buckendahl, Cindy Gray, Bob Henson, Jeff Nellhaus, Linda Poole

Present: Melinda Montgomery, Steven Courtney, Aly Martinez-Wilkinson, Chris Meador, David Cosio, Mayuko Simon, Jing Chen, Katrina Fitzpatrick, Jeremy Heneger, Lee McKenna, Allyson Olson, Trudy Clark, Stacey Larsen, Sharon Heater, Christina Schneider, Garron Gianopulos, Kara Courtney

April 19, 2021 TAC Minutes – no corrections, approved as presented

Through Year Model Simulation Results
During the NDE, NWEA and TAC through-year test model working sessions, the NWEA psychometrics team received feedback on configurations that stakeholders would like applied to test models for simulations so that the test design will match the desired outcomes. Based on the feedback, the current model requires the engine to limit the selection of diagnostic off-grade items to one grade above and below the student's enrolled grade of record. Additionally, the engine focuses on information in the diagnostic section from domain areas where the simulation shows specific areas of weakness. High-performing students are simulated going off grade when the result from the operational section illustrates high performance. NWEA will present the results from the updated simulations for ELA and Mathematics Grades 4, 5, and 6. These results will include the test reliability, blueprint matching performance, diagnostic section results, and examples of scoring options based on the simulations.

1. Based on these results, what additional updates would you recommend in preparation for the Winter Pilot?

TAC: Define diagnostic growth – on and off grade growth items; tech suggest 65% of items are delivered on grade (as first 27), but table shows higher than that – is 65% underestimating the actual distribution? No, it is within the parameters. Those averages are across the population that took test in simulation, yes.
NWEA: Limited to one grade above and one grade below. No decisions yet made about going more than one grade. After linking study will run additional simulations showing this.

TAC:

- ELA will choose a passage set below grade level? Yes, may get different P value - may not get all items from passage, but will get several items for passage before sending them to another area. Have set a minimum of 4 items per passage, for both MAP Growth & NSCAS. Does MAP G have similar constraint regarding minimum number of items per passage, or is this individual standalone items? Yes, more standalone. They have both standalone & one with 3 items.
- Reason ask this, because a lot of districts using MAPG, if dramatically different from what they've experienced will create anxiety. From policy aspect need to follow similar pattern of configuration. Is it a timed test? No.
- Last time talked about ceilings & floors: implication of ceiling on a MAP score, but what is implication on a RIT score? Concerned with parent will see high percentile rank vs lower RIT score. Preliminary at this point will look to compare the linked RIT in relation to MAPG (Does it actually cover the score range) – Implication of a ceiling on a RIT Score? Does it work the same

are percentile rank? In MAP Bank, 2 grade ranges 3-5 & 6+, so have ceiling in math: highest scores are ranked order with cumulative frequency distribution, that is used to develop percentile, but this model is bringing in MAP flavor (i.e., see items grade 3 & 5 for grade 4 test)

Scoring Discussion: Scale & RIT scores from Winter Simulation data: What is correlation of ability estimate between 27 & 40 items – NWEA did not calculate this; curious – idea if give more items then estimate of ability becomes more reliable. Even if go off grade may actually be on grade. Whenever go off grade, difficulty of items calculated on same scale, may be closer to personal theta. Struggling with weighing two scales. NWEA: True theta to final theta correlation between 40 and 27 items?
TAC:

- Actually 3 correlations: true w/ first 27, True with the 40, and 27 and the 40. RASCH model suggests more items, as long as item parameters on the same scale, means more reliable measure of ability. Will need to rely on the blueprint. Good start as long as people can make connection between MAPG familiar with and system reasonably linked to it, you can make transition. Policy-wise, what is flexibility to make changes if find larger samples and full item banks make a difference?

NDE: yes, believe we can make modifications before making full implementation – minor tweaks to constraint engine, for example.
Additional analysis: if strong relationship between 27 & 40 item score, because accountability system based on student classification, may have smaller standard errors around consistency values = may have greater confidence if use 40 items because can go off grade. What is risk if student changes category, (since do not go into diagnostic until decision made after first 27), but from reporting standpoint, don't lose anything using 40 items to make final determination on score – greater confidence perhaps.
When using 40 items, is blueprint maintained? No, can go outside blueprint (only at grade level) – may have problems with peer review due to this.

NDE: having more evidence, greater reliability & less error should be good thing; in some ways will be a philosophical argument vs psychometric argument; classification will always be based on on-grade level items – never in a position where students classified as college & career ready will obtain that designation because of off-grade items. Seems to be how you make the argument to USDOE, because decisions made based on grade, monitoring potential for student moving categories in last 13 items is important. Could be problematic from accountability standpoint but doesn't change Met/Not Met decision point.
NWEA: look at scale score for 1st 27 items & cut score then scale score for total 40 item test & compute percent of students who stay in same category or shift to another- can do this with simulation have now but will do it in the full CBE. Will want to look at it with better matching on the blueprint.
TAC:

- Sounds like potentially basing scale score & RIT on all 40 items, but the classification on first 27.
- What score will be associated w/ classification? Will there be two scale scores? Yes, will be two different scale scores
- Looking at score report for parent and school will be helpful – work on how to communicate data – be familiar but also better.

Item Difficulty Modeling for ELA Reading Items 57 min
Item difficulty modeling (IDM) was conducted to predict item difficulty for NSCAS ELA reading items using item features that reflect cognitive, content, and stimulus demands of the item, including item type, item RALD level, DOK levels, and passage text complexity. Four models were applied, including the

widely used multiple linear regression and three machine learning algorithms: support vector machines (SVM), random forests (RF), and k-nearest neighbor regression (k-NN). Results from the models suggest that the included item features can predict item difficulty to some extent. The linear regression model and the random forests model achieved an R-squared around 0.38 and 0.41 for the reading items, respectively. A moderate relation between item RALD level and item difficulty was discovered, which provided validity evidence of the RALDs and the items.

TAC:

- How was range ALD and DOK determined originally? Content specialists developed them Previously established, correct? It wasn't part of NWEA's study? No, was in item developing process. Writers assigned with alignment study with content specialists.
- As a predictor using range ALD and item difficulty, .39 doesn't seem strong predictor in explaining variance. Could you explain the interpretation of data & what thresholds using (i.e., good, moderate, strong)?

Jing: IDM literature see wide range of .1 to .8 so range ALD correlation point .4 as individual predictor – feel is promising. Would TAC have criteria?

TAC:

- Feels like needs to be higher – explaining 16% variance seems modest in terms of predictive. If incremental added value of predictor and there are other indicators that adds value ok. But not as a standalone predictor need stronger. With multiple regression will see overall prediction. What's good enough depends on what you are doing with the data. To what extent this model used to help write items. If not high stakes, can Inform but not the only thing that informs. Item type has influence as well so if looking to differentiate at higher performance levels and introducing beyond multiple choice seems important. Why some TEIs more difficult is another question – What made TEI's more difficult? Student familiarity vs content? More constructed response so more authentic measure? **Christina** - in literature what see is that more steps that are required in solving a problem then see greater difficulty so TEI help with that. We are getting some truncation when look at percentage of items at higher DOK and the ALD.

TAC:

- Goes back to what standard is calling for – multiple step processes needed so more difficult. Range ALD are based on standards? **Christina**: Range ALD intended to be but also talk about intersections of students with what asked to do for a task and how author setting up moment to retrieve detail from text to conduct an analysis or make interpretations. Looking at how author sets up a scenario or gives context clue. Can see these nuances in different passages. Becomes a progression of when is child able to demonstrate skill based on the passage. Intent of the RALD

TAC:

- Why is Lexile (difficulty) not correlating with item difficulty? **NWEA**: yes, is showing that pattern. Lexile range between 0-1500; if improves by 100 then item difficulty increases to some extent. **TAC**: would guess play larger role than it appears. **NWEA**: Context and interaction determine difficulty more directly. A debate in reading community about role of text complexity in reading comprehension vs. role of task of inference and conducting analyses. What see is that within passage have items that will target students at different levels of reading development. Seeing here grade level is not driving difficulty, but the Lexile and number of words are. Jing took out text complexity measures and ran a model based on this and grade

became predictor when text complexity not there. If look at impact of coefficient on 100 points on Lexile scale and see movement from 400 to 600 then see larger influence with complexity. Not capturing all pieces of text complexity in these models. See larger impact on item parameters.

TAC:

- Within models, focus seems to be on text complexity and sometimes type of text – literary/informational, allegory, poems – linguists get into text difficulty without considering interaction with task. Plausibility of distractors in multiple choice is not considered. The more difficult passage could ask literal comprehension question vs. having a simple passage that is inferential. Depending on the distractors, we may be missing key piece of item and information associated with it. This is missing in IDM we're talking about. **Christina**: This is exactly right. Distractors and complexity of stem are difficulty drivers. Are considering looking at the complexity of the stem, you can get difficult passage and literal retrieval question making it easy. Want range ALDs to exemplify authentic experiences and not just what's in multiple choice test. If we are interested in trying to understand cognitive processes of reading, then am philosophically opposed to put distractors in there just to increase predictions.

TAC:

- What metric would you use? **Christina**: looking at complexity of stem is good. Like to get better at writing the range ALDS. Have more constructive response items - will also see writing and ability to get kids to parse the text, then can start seeing kids take 2 passages, do authentic tasks, and read and write. Realize is a lot of scaffolding. Interested in getting more authentic model. **TAC**: Students at different ability will interact at different level. If building item sets with passages with multiple items associated with it – different approaches to levels and how interact with. **Christina**: support idea of task, text complexity and reader – we can't model the reader (maybe through engagement & interest statistics) but can model the complexity.

TAC: Regarding the table (coefficients of independent variables), if had some kind of measure of effect size than easier to talk about importance of relative performance beyond significance. Don't know what is more important. Effect size would help.

1. What are ways to communicate results to teachers about the value of RALDs given that the RALDs appear to contribute unique information to the model and are barely to not related to text complexity?

TAC:

- From field perspective to what extent are teachers using the ALD? Experience indicates don't use them much – going to standards more. Is important to familiarize teachers with ALDs. **NDE**: Trying to make them more useful to teachers. Trying to help understand how students can move through ALDs. Used not as much as want but will increase moving forward. Linda: agrees teachers looking more at standards. Cindy: perception of some educators - feels ALD is teaching to the test. Teachers are experiencing standards overload - can't keep up with changes to standards along with another document. **NDE**: Working on RALD tool. Meant to give back info in same way as learning continuum and can look at as a progression of learning. **NWEA**: Provided ALD PL to take items have and tweak to differentiate instruction and adapt what already have. Have content advisory boards. Conclusion: if educators are aware and know how to use, find it valuable. **Cindy**: Teachers fearful that NSCAS Growth won't have same info as

MAP (the learning continuum). **NDE**: Won't go away – improving it. **Christina**: need TAC help teachers see difference; descriptors are same across whole grade level – want teachers to see these descriptors give context that child able to infer theme and main idea when not explicitly given and had to use implicit information to draw conclusion. Trying to get to understanding when a child is putting cognitive pieces together - when child can do it and what's happening cognitively with child to get them there.

2. How should the range of text complexity be considered in a through-year system? In supplemental analyses, we have found that grade becomes a significant predictor if we remove text complexity. However, with Lexile in the model, Lexile becomes a surrogate for grade level even with the wide overlap of Lexile levels across grades.

TAC: Text complexity influences but doesn't determine difficulty – this is the right message. Extent to which it contributes to modeling is better - The better models are to predict difficulty more can do for embedded standard setting. Should be included, but how included, there isn't consensus.

Embedded ID Matching: ESS Enhancements to ID Matching to Reduce Panelist's Cognitive Load
EIDM Standard Setting Method
The Embedded Item Descriptor Matching (EIDM) method is intended to integrate alignment and standard setting into a single process, taking what is often viewed as disparate processes of assessment development and synergizing them. Under EIDM, teachers align items to range achievement level descriptors (RALDs), which explicate the progression of cognition students need to show for more advanced knowledge and skills in the content area as they develop toward college and career readiness or other appropriate goals. Rater agreement statistics quantify the degree to which teachers agree on and converge in agreement on alignment across each round. The Embedded Standard Setting (ESS) method (Lewis & Cook, 2020) is used to calculate the cut scores based on the alignment information by optimizing the placement of the cut score at the point on the test scale where score interpretations are most accurate. Therefore, this approach offers states critical advantages over treating alignment and standard setting as separate approaches; it adds value to test scores while streamlining processes.

1. What is the TAC's perspective of piloting EIDM with a small number of teachers or NDE SMEs in science with the field test items and adjusting the RALDs based on data?

TAC: Seems that this method depends on range ALDs being very comprehensive. Don't have tasks reflective in ALDs on the test. In past ALDs was more representative of standards. **NWEA** – wouldn't use process unless a principled approach to test design.

TAC: Sounds like they never see impact data about what students end up in category. Had cuts at top end and labelling. Teachers thinking have to be above 90 percentile to be college or career ready? Is there a place to see impact? **Christina** – impact data is requirement of standard setting. Teachers are focusing on alignment and policy makers focus on policy review. Key stakeholders focusing on alignment and setting cut scores based on expertise in content.

TAC: Would support moving forward with pilot study to evaluate and refine methodology. This is Deja Vue from 1996 with bookmark when standards set, and policy definitions were then changed. Disagree with this approach. Range ALDs is meant to be policy guidance for standard setting committee. Should be state driving that. Because data dependent model important to be confident to where items map to.

There is overlap with range ALDs mapped with item difficulty. Policy responsibility should be on policy body. Outcome becomes policy tolerance activity as step to what come up with recommendations. Get cut scores set up by design up front but don't underestimate what policy makers decide. When policy makers adjust then the Department has to adjust the Range ALDs. Like the idea but would say is a policy driven process, so don't mess with ALDs after content. Wait for policy makers to interject and then revisit. Takes responsibility away from teachers who should not be doing this work. **Christina**: rather than adjust ALDs in third round, make sure alignments super tight and calculate cut score. Focus on alignment will get better. After policy, make adjustments if have to at that time. Rationale based on policy on why adjusting ALDs.

TAC: Less than if but when. Never seen policy body – rare adopt it without feedback especially if teachers didn't have impact data. Also an issue of moderation across grade level. Practically speaking when will that happen. When cuts go to policy need to have a range recommendation. Process has integrity and a percentage of people in agreement, but possibility of moving up and down. One of ways to characterize it is more explicit judgment activity but after analysis & recommendation phase then it is like contrasting groups. Trying to maximize classification consistency in finding point of distinction representative of where one ends, and another begins. Nice blend of content anchor to minimize misclassification.

TAC: Is there anything about science in particular that concerns you relative to applying this to math and ELA moving forward? If results are successful. **Christina**: item bank does not fully represent intended blueprint based on what seeing on alignment. In terms of intended content pieces that department prioritized as wanting to measure. One of grades has 2 items that are classified at high ALD level and Department has moved to use ALD as the cognitive model. Know that have to be very strategic and worked hard to build templates so we can build tasks this summer to target areas of blueprint that need to fill. **TAC**: Item clusters – up to 6 questions? **NDE**: yes, tasks and prompts within, are item clusters. **TAC**: How do you handle items harder in cluster. Doing pilot? **NDE**: yes, take this approach with field test done and can see if move forward. **TAC**: Wondering with pilot have 2 groups at grade level have variation in procedure to see if some works better. Opportunity to do NAEP-like research. How many people on panel? **Christina:** Doing 3-6 people on panel. Content team has tagged bank with ALDs so have those as well. Other considerations re: Science assessment.

TAC: That sampling element – similar to ELA and clustering standpoint. With ELA, a lot of skills are threaded throughout but science isn't the same. To Jeff's point maybe this is a research opportunity – if could have two configurations that represent a different sampling (context/content) to see if groups converge could be useful experiment. Whether for ID matching or any standard setting, test must have capacity to support number of cuts trying to make. If don't have enough at upper end to do second cut, then focus study on cut know have distribution for and do met or not met to avoid results that suggest there is a problem.

TAC: Always concerned with science because not continuity with concepts taught at grade levels across districts. Does assessment match the level of exposure of tasks in class vs. the progression of knowledge?

TAC: Does anyone have concerns or problems if Department moved forward with piloting methodology? If this is a new method and criticism would be: would data come out differently if using traditional/bookmark method; if piloting would this method produce wildly different than this model? People may say it is method not the student. Think through unintended consequences. On face love the method but think of implications in the field. Only dealing with 2 grade levels and if could do control group for point of comparison might be good suggestion. **NDE**: We could do one group with this method and another group with traditional. **TAC**: if have grad school student needing a dissertation this could be good research.

TAC: Question is what is purpose of pilot – try out procedure or something else? What is the real question – get best results, implementation? **NDE**: ALD info more useful to teachers is some of it. How do assessments meet the purpose of the law and make it more useful to teachers? **TAC**: what is the metric? How do you know success? **Christina**: Rater statistics & reliability is important. Hypothesis is Range ALDs improve coherence in educational system. Checking rater agreement is evidence that teachers can do this. For pilot, look at how good is agreement can get in terms of convergence? We want to see if teachers have agreed but do not agree with cuts following completion of assessment. How are we moderating it, what are you going to do with the policy panel, will have to re-tag it so we can get an assessment that provides two for one.