Nebraska Technical Advisory Committee Meeting

Nebraska Department of Education

April 19, 2021

Present: Chad Buckendahl, Bob Henson, Cindy Gray, Jeff Nelhaus, Linda Poole

Also in attendance: Stacey Weber, Aly Martinez Wilkinson, Jeremy Heneger, Christian Schiller *DRC, David Cosio NWEA, Jin Chen, Kara Courtney DRC, Kelly Manning DEED Assessments Admin, Lee McKenna dRC, Maggie Sis, Christina Schneider, Mary Veazey, Mayuko Simon, Melinda Montgomery, Rhonda True, Sharon Heater, Shavonne Jordan (Alaska), Steven Courtney DRC, Hongwook Suh, Deb Frison, Allyson Olson, Iris Owens, Christopher Chambers, Karen Barton (NWEA), Katrina Fitzpatrick, Lane Carr, Lisa Fricke, Shaundra Sand

Jeremy began by reporting on status of testing. The TAC members were introduced.

Chad Buckendahl, Chair, called for the minutes to be approved: Jeff Nellhaus moved, 2nd by Bob Henson

Using Principled Alignment and Item Difficulty Modeling to Create an Integrated, Hybrid Interim-Summative Through Year Adaptive Assessment System: Proof of Concept and Considerations NWEA presented on building a coherent assessment system with merged constructs for the through-year assessment. NWEA stated the Range Achievement Level Descriptors (RALD) would be the bridge between the two constructs of interim & summative assessment systems.

TAC:
Does this require a much higher item pool than normal since in a given standard there could be a range of complexity? Yes, so trying to build much larger item pool than one finds in a traditional CAT. If want to measure progression of learning, then yes to meet accountability targets as well. At min, based on length of assessment, need 800 items if pool meets specifications.
Theoretical strategy NWEA working on – bridging what we love about interim & summative assessments – will meet through item difficulty modeling – identify task features expected to predict item difficulty along a test scale. Predict item difficulty using the item features through a linear regression model. Document the performance of the model in predicting empirical item difficulty.

NWEA shared 3 studies: 1) Is RALD a good predictor? 2) Determine if summative & interim assessments could be linked to develop coherent inferences regarding student learning. 3) Determine if RALD-to-task matches to the item difficulty model could be used to determine if interim test items could be optimized to support a single, integrated approach of connecting summative & interim purposes.

Until NE updates standards, we will want to keep characteristics the same as much as possible so NE will not need to go back to peer review.

TAC:

Summative is NSCAS test correct, yes. It does look like for 5th gr in both tables that seemed to be particularly challenging – 42%/40% - any thoughts around this & 7th grade (NWEA – either item is challenging, or raters have error – need to review data more closely; tension in summer in doing this quickly and not being able to work with raters who are having troubles – recommendations for NWEA?) Related question from TAC - were these 6 separate panels or in grade bands – (NWEA per separate panel). Sometimes standards & ALDs at 5th grade because it is the end of elementary & transition to MS, may have something to do with it. Was there something systematic about adjacent ALDs? Patterns like were panelists going to ALD above or below in adjacent agreement? Whenever you say it's being investigated, what are you looking at? Perhaps it's the independent variables or perhaps multiple raters are causing the problem (inconsistencies). Perhaps raters can still be reasonable to use if more training.

- Additional item types used, are those largely used to get to higher levels of DOK? Are you surprised that multiple choice (MC) had negative coefficient? Aren't other item types expected to raise item difficulty. General misnomer that MC are easier or cannot get to higher order thinking skills. This is part of design, if other item types not there, not contrast for item types to be predictors; wouldn't want message to be MC makes items easy. Jeremy – help us to show range of difficulty of MC items? Chad – particularly useful to do this. Rest of TAC, yes helpful – if reporting out these weights easy to over-interpret results – provide marginal tables will soften impact –
    - Use regression – TAC depends on group – too much in weeds will I need to get even further vs global – be careful of interpreting weights – marginal tables may be useful; most people at districts will not dig down; inter-rater reliability is more important – are we only agreeing on 3 decision points? (Cindy) Perspective of impact on targeted grade on item difficulty vs RALD? For summative grade level but not on average as much as RALD, but Interim doesn't have same relationship – things to consider going on- off grade; Cindy- have more questions about sufficiency for decision making (number of items off grade level) + level of error of this relationship
    - Wondering if construct is different enough between 2? NWEA – hot debate in field if whether best to stay on grade or off grade for students; measure growth vs what need to know (social justice) – within this paradigm does the data speak to these interpretations? Bob – results do not speak to this paradigm, rather the constructs seem to say they are about the same; targeted gr vs not: evidence not quite the same – doesn't speak to whether it allows to go off grade- effect size could help identify this, but may be a small significance

TAC

RALD to Task Match table – items in pool - not enough data to support decisions at higher levels/ top performing students; NWEA – not enough items at Levels 3 to fill gap; field test hopes to target these items; accurate for data, but will improve; recognize this only showing gr 3 but can we extrapolate to other grade levels- field looking at inter-rater agreement & raising questions – most items targeting

Levels 1 & 2, these agreement levels more concerning since no expectation of uniform distribution of agreement – opportunities for disagreement wasn't there with only 4 cells (Level 1 & 2 and DOK 1 & 2); need further item development

TAC Guidance:

- Outliers, before excluding, think of other variables could collect – spend some time identifying reasons behind outliers; not able to find it or other reasons why, then maybe exclude them; be cautious to exclude things from analysis, if just a few items might be worthwhile
- Really neat way to compare different assessments; modeling item difficulty based on something else & then bringing interim to summative is neat idea
- Is ultimate plan after bringing some of the items over, continue to pilot them or will use predictive item parameter? Use calibration from NSCAS scale
- A different study: Would be interesting to see how different the theta estimate when use predictive item difficulties vs the predictive item difficulties – might find predictive model does reasonably well (sidebar) - easier way to interpret .45 for the R-squared – depends on setting, but may not change your theta estimate
- Tying item types to item specifics for RALDs – whether you want to identify types based on evidence statement implying – not bad idea; should be something implied in evidence statement that drives the item type; different item types used randomly rather than evidence statement drive best way to measure evidence – evidence statement might suggest more constructed response but need to take clue from standard or evidence statement – help item writers & especially getting questions at higher levels
- Connecting systems – is it talking about different constructs or features that influence? If talking about different constructs, worries about dimensionality concern, this is different question, but if asking how best to incorporate influencing features, Jeff's suggestion is one way to another question of item specification; communication to go out may be no expectation of uniform distribution across the scale – not many in field is familiar with this – If try to rely heavily on content specification, will lose efficiency of the CAT. It is important to communicate how this testing experience will feel.
- Conversation about peer review
    - Model robust enough to support this sufficient measurement approach, can demonstrate alignment of item bank to standards, and the personalized student experience to maximize efficiency to ability of student
    - Worst may say is to put on content constraint
    - Force them to reject it – now more support broadly at peer review level
    - Peer reviewers get caught up on content constraint (5+ items) that creates challenge even within grade for representation – Joe Ryan showed with content clustered with graphics that content still there on scale

Through Year Prioritization Model Overview – The prioritization model delivers a set of items based on the on-grade blueprint at the beginning of each test administration. Based on the student ability classification from these initial items, the blueprint adjusts on the fly to maximize what we can learn about each student. Students who are proficient on the content delivered in the first part of the test will have access to more complex on- and off-grade content. Students who are classified as diagnostic will receive a narrowed blueprint around the areas of the content where the student is the weakest.

TAC:

When talking about this assessment is it during the year or summative? 3 x a year – when say adjust the blueprint? What does this mean? Won't measure full range of blueprint or narrow it? First 25 items are proportional to current NE blueprint. After these we can tell where student is & 2nd part of the test will go beyond. Blueprint covers all 35 items will be covered – Is it like a staged adaptive approach? No, the blueprint will be narrowed? Does it narrow based on concepts? Yes How is this different from MAP Growth? This is on NSCAS item pool/aligned to summative blueprint – targeted to NSCAS blueprint but allows flexibility; First 25 items are adaptive? Each student different. If have previous information can determine substance of test. In CAT, 10-15 items will be enough to figure out where a student is. First 25 Q are adaptive, not fixed form, & can start w/o previous information. It will cover full range of blueprint – not narrow. NWEA maintains the questions will be proportional to blueprint. It is adaptive based on Rasch? Yes.

Inconsistency about naming proficiency level – using College & Career Ready is hard for parents to interpret. Naming is problematic. What information do we give teachers? If give information in the fall, does this have unintended consequences? What information do we present? Do we label kids before the end of the year? We want to keep emphasizing that these are interim and want to be careful with summative info. Are we thinking we give the same info such as growth projections? Some may think need to use MAP Growth still. Jeremy: intent that will serve both so give the same info - no double testing. In many cases will be redundant. How long transition period? NDE: next year is partial implementation and 22-23 will be full implementation. Cindy: if RIT scores are significantly different from MAP Growth to NSCAS then it will prolong the transition. NDE: we will get a RIT score based on linking study and how different they are matters. Will be opportunities for comparison and that will be when districts make decisions.

TAC:

Is purpose of different direction (challenging/less challenging) to increase precision of lower or higher score on scale? Is this really diagnostic information? Purpose is mechanism for increasing precision of item? NWEA - Want to give feedback to particular ALD. Is providing information on what a student can do – if cannot do 4th gr can do 3rd grade in concrete terms. When we come back, it would be good to look at the score reports. Then we can see what we are getting – we can work back & show how we get there.

TAC:

Diagnostic items – how are they selected? They are adaptive selected on where student's level of proficiency is and then chosen by item difficulty level. Are challenging items also adaptive? How are challenging items & diagnostic items different from other items? Focus on content decisions/constraints change. Looks like after first 25, if are at developing then you stay at developing? No, still continues being adaptive. Not just diagnostic, but also feeding the final designation? Will be a scoring decision. NDE: When we give summative determinations is the piece we have to be very careful with. This graphic is a simulation and want to be careful about how people feel about interim. This is only simulation – all things could change.

TAC:

Do we have a sense of distribution of difficulty on the scale? Are items selected relative to student ability. Recognize content areas are not scaled to the same part, it will be more difficult to find some in the moderate and more difficult with current bank. We think about blueprints as proportional representation based on fixed-form models. Is it more appropriate for adaptive to say we have 4 sub domains interested in measuring and will have a minimum of certain percentage. Start with assumption may have unified distribution but based on where student goes, may have more of some content than others. If subdomains are not uniformly distributed in item difficulty, we may lose some precision when try to fit a traditional "blueprint" mindset. As do simulations, could think about playing with relaxed constraints and number of items to allow for greater precision and getting to personalization.

Classification decision- think peer reviewers concern is when talk about when on or off grade. To make determinations, if have evidence to support that at a particular point in time (a student meeting end of course requirement) – less grade dependent and more at ability level. Will have evidence to make the summative decision & replicate this into the Spring. Policy business rules of how function, once make summative decision can bank that and rest is a formative/interim purpose even in latter part of year. Becomes a data management issue. NDE: questions around banking proficiency will have to tackle. Could possibly bank it over years. If not in top level, can they still increase based on evidence that comes back. Focus for peer review would be when do we have enough evidence? How much evidence do we need before go off of grade level? As talking with stakeholders, to make any summative decision before winter is problematic because disconnect the assessment experience from curriculum and instruction. If the decision to drop below grade comes too soon and haven't provided instruction to have allow student to catch up, run into social justice issues.

TAC:
You've mentioned subdomains. Want to be conscientious of unidimensional model & using an adaptive test based on this model – the two approaches ignore blueprint. May be difficult to get distinct info on subdomains. May get less reliable measures of the unidimensional model Can think of subdomains you are selecting items for, but model sees items only as item difficulty - may still be getting only half of those items right as opposed to the actual domain. How the items selected & how the distribution of the difficulties look could actually determine how you report back rather than where they actually are.

TAC:

You are doing this with only 35 items and the feedback other than on the overall domain is pretty limited. If give feedback that the student could use more instruction in a subdomain, how reliable is this feedback going to be with the number of questions you have.
Regarding content decisions: If Nebraska content people are comfortable with those prioritizations and decision rules then those are content decisions.

TAC:
Seems like proficiency score should be based on all info can gather. Why just base it on the 25 or additional questions? NWEA: Noticed that so far that score can go up or down after 25 questions based on items receive. How we handle students at the tail is more the concern area. TAC: How you use it is the question. When simulating is it true theta? Yes. When theta goes up or down is it closer than when use 35 vs 25? Ask because when you use the 35, may be bias should be less and less biased the more number give & should have smaller error. NWEA: varies based on where student went off grade level during diagnostic section. Diagnostic section is zeroing in on one or 2 domains, so if think about proficiency scores, with oversampling 1 or 2 domains be comparable when sampling the whole blueprint. TAC: In simulation do you simulate domain scores? It may be honing into domain but if simulate single theta & simulate responses from the theta, domain is a language thing. NWEA: It does when it changes to part one vs part 2 it does change the focus of the blueprint. TAC: when simulate responses to domain in second section still on theta and item difficulty? NWEA: yes, not just item difficulty. TAC: simulated response (yes) You can say items are narrowed down but if still only depends on theta & item difficulty. NWEA: Doesn't just depend on item difficulty also depends on identified content area. Weighted penalty model. TAC: Simulated response is now function of the difference of the difficulty, theta, & content. There is going to be reality vs model. In simulation, if domains exist in simulation to see what happens, it might be worth it to include a model misfit (introduction of actual domains in addition to generalbility) – will show you closer to reality.

NWEA and NDE would be interested in having working sessions with TAC members – Bob, Chad, and Cindy are interested.

TAC:

Difficult to explain proficiency levels drop when questions are out of grade level. NDE: Not sure we will make that argument. Grade level proficiency will be based on grade level items. The growth scores might reflect off grade levels. TAC: RIT may be but not proficiency levels. Peer review have problem with using off grade level to determine proficiency but may have issue with growth too if it feeds into accountability level and how strong that claim is. NWEA: How do you interpret when there's an allowance for the assessment to adapt above and below one grade but still provide on grade proficiency? Do they score only on designated grade items? TAC: Yes, to know how to build out the argument would like to find out what Smarter Balanced states are doing with that and look at accountability systems if doing anything with off grade level content to inform growth. Anything beyond that is considered for instructional purposes only. NWEA did research for SB schools: Assessments have to have 62% content delivered on grade. If child on top or lower achievement level then can go off

grade. In accountability systems, have summative determination for proficiency is calculated, but total theta goes into estimating growth.

TAC:

In terms of NSCAS scale range, to score at top of scale do you have to answer off grade level items? NDE: Not currently but with through year students can go off grade level and we would be able to capture growth. TAC: The off-grade level questions may impact the RIT scale score but not NSCAS scale score. NDE: Depends on how will break it up. NWEA: Want to caution that there are a lot of purposes trying to hit with any of the models. Making sure we are balancing instructional value information that teachers & students need throughout the year, which requires complexity and off grade items to personalize where the student is & inform instruction and show growth. MAP Growth is completely grade agnostic diagnostic. Can we get enough information about level of proficiency, to continue to evaluate maintenance of that knowledge through the year? There is a tension between assessments not taking a long time, provide valuable instructional information, and do we need to ask same kind of questions if the student is proficient. These are policy questions that come down to business rules in adaptive environment. Is peer review the ultimate arbiter of what is growth, achievement, proficiency, how do we provide instructional information, or not. That is where we could use the TACs guidance; for NDE from a policy perspective, how do we navigate this kind of innovation for their state to hit all of those priorities and not mislead instruction. TAC: Related issue is standard setting in time of pandemic. Students who are not participating is a non-random group so setting proficiency levels based on a less than full population and the part of the population that is missing is non-random. NDE: We're not setting standards on this year results. TAC: On the standard setting question, unless state is compelled to do standard setting, other TACS in other states are saying don't do it. Nonrandom section of population may be over or under estimating performance so 2022 is better for standard setting.

TAC:

What metric for MAP Growth? NWEA: Uses percentile rank, growth is likelihood of staying in same percentile rank across time. Depending on where you are, work on increasing over system projections with growth of the child. TAC: Have to be careful talking about growth percentile. Difficult to understand difference between percent and percentile. Interpreting is difficult. Important to understand the metric and interpretations with MAP Growth. In terms of new assessment system, what info will go into the metric? When set goals it is helpful to see how many RIT points must go up. MAP is helpful to give graphs to see where can close gap. Question is what do you need to do to improve it score? What is required in terms of learning? NWEA: Hope with ALDs you can see cut scores. If growing but still hanging out in novice state of content area. Idea is with achievement level descriptors offer more nuanced info, then teachers can say where have to grow in content pieces, then have a score and content to grow in. This is what is valuable to teacher. Disengagement tool is popular– want to see that stay with NSCAS. NDE: Not sure where we are with that.

1:45 – 2:15 Additional Discussion Through-Year Model

Considering a Through Year Assessment System (nciea.org) By Dr. Brian Gong

TAC:
Thoughts of Brian Gong (on NE TAC previously) – for future meeting as flesh out ideas on business rules/decision making & how it plays out – incentivize appropriate behaviors & discourage inappropriate behaviors & how TAC can partner with NDE.
Semantics and what call model is important with this. Through year, balanced assessment program, or prioritization model. Matter of semantics – what to call it so it is not misinterpreted. NWEA: Have had conversations with Brian on informal levels. Think is important to get feedback from the field. Want to provide most valuable information in terms of outcomes. Better to hear from others about the process.

TAC:
More you can visualize what score reports look like, what new information be reported vs what is currently reported, then can evaluate the strength of the information (where it is coming from & technical challenges of providing it, & the options of providing it). Know where going. Better can define it, the better we can help you make those decisions in terms of the methodology. When get closer to what model look like, useful activity is using focus groups with score reports put out and asked how they interpret it to see if made correct interpretations. Use strawman models. How would field interpret. NDE: NWEA is working with feedback on score reports. TAC: If RIT score influenced by additional items but proficiency level is not, people will notice if students with the same RIT score end up at different proficiency levels. Hurt the credibility of the assessment.

2:15– 2:30 Next Meeting and Adjournment

May 27 from 12-4. Will get additional info on working sessions from NWEA

Adjourn at 2:21 p.m.