Nebraska Technical Advisory Committee Meeting
Nebraska Department of Education
November 18, 2020

Virtual Meeting
11:00 am - 2:00 pm Central Time

TAC Committee in attendance: Jeff Nellhaus, Bob Henson, Chad Buckendahl, Cindy Gray, Linda Poole (11:30)

Opening remarks: Chad – few more frequent mtgs to try to stay abreast of topic areas particular with TY assessment – want major discussion on TY research plan – novel plan & want NE to be ready to carry it out & have data to make good decisions
Introductions

Minutes approved

Spring 2021 Alternative Test Design – Follow-up Analyses

- Spring results are not the same since they are horizontal equating -
- TAC – call out what is specific to 2021 and for TY design
- NWEA
  - Report out in content area only for 2021
  - Simulation on classification accuracy – relatively good – no place really off
  - Issues at reporting category level – not quite as well at distinguishing – reason why suggest reporting out in content area
  - Subgroups – variable in engine – engine wasn't biased – will need to check on early returns
- TAC:
  - How did you previously report sub scores (metric used) – used NSCAS scale not RIT - used level of achievement but had 10 items not 4– had both achievement level & scale score? Both with 41 items
  - Only this year (2021) will report only in content area? Yes.
  - For TY will report much like a MAP Growth, 3x a year
    - 2021 test will only tell us if we held our own
    - Not comparable, no formative data
  - If reporting on RIT scale, goal is to give TY experience so can compare fall/winter
  - Will it be comparable for high achieving students? Not enough of a ceiling for comparison?
  - NWEA- distribution should allow comparison
  - Scoring – on RASCH model
    - In support not reporting any subcategories
    - RASCH will put all on same scale, but if fits ok
    - Subscale will be less reliable if assessment is unidimensional
    - Subscore analysis: engine worked well & no indicator of bias
      - Subgroup more around performance & more of DIF
      - Will Sp 2021 give any indication for hypothesized learning loss
      - Equating strategies from TAC
      - Jeremy – overall part of what will review – RIT score piece, but other measures will do most of COVID research
      - State does not have statewide/national data – 65% of normal testing took place – saw some loss
      - District level – fall testing but concern was some students did remote testing – no analysis – results questionable – Fall results are of limited use

- Reliability – no accountability for this year due to uncertainty of reliability
- Reporting categories: given no comparison, will provide some normative information at school level to provide contextual information? Have you thought of other ways to report out subscores? Raw data? Compare school vs district?
    - Jeremy – how much & ways to report out at State level? Want to give as much info as possible
    - NWEA – consider as fully adaptive CAT – while only 4 items, so no student may see same items as another – because of this may not be able to give district vs state level information
- If did report subscales as average, way to increase "precision" – may not has as much reliability for individual but average may have it – test it out, as it becomes more reliable – if hit level acceptable, can end up giving schools similar scores across subscale
- State level information – create peer groups (hybrid vs fully virtual vs all in school) –
    - Is this captured? Other states have considered
        - NDE is trying to capture this info – targeted data from schools at student level (days per model) – varies a great deal due to pandemic and quarantines
    - May be only an interesting research question, but is that type of information potentially useful for districts or only useful at school level? Good to have factual data to address parents/union concerns
    - Could do a secondary analysis this summer – collect instructional approaches from schools – will create rich research opportunity
    - On right track to identify research questions so know what data to capture – important variable is how many remote students take the 2021 test
- Linking/equating question – recommendations in reporting scores based on linking studies - is equipercentile equating appropriate for Sp 2021 only? Did other analysis to see if common-person linking would work under different situations – equipercentile seemed to be better distribution
    - Concerned with score shifts under certain conditions
    - RIT score in spring will be more useful
    - Going into TY spring after 2021 want common item linking to put on same scale
    - TAC: using same students taking assessment fall & spring or different year? Ability distribution for which year? equipercentile mapping will use 2019 data if took MAP
        - Using equipercentile, there's an assumption of random equivalent distribution of population, but will change in learning environment violate this assumption – masking differences in distribution
            - Comparison was NE students to NE students – but will look at fall/winter distributions to see what adjustments need to be made
            - Given not using test results to compare to previous administrations perhaps less of an issue
            - Isolated assessment – standalone test – look for ways to identify cutscores through scale
            - Not sure of a better alternative – result in some useful information since will not use equipercentile going forward
    - Risks to use this information for learning loss? If methodology applied, are any negatives for this? Any skepticism from field or public? Statistically appropriate

way to transition from old to new, but any negative consequence using this strategy if there is another strategy that would show true decline?

- No public reporting of test data
- No building/school reporting, but districts will get data
  - Concern- once you gave individual student scores to district, media will ask. Only provide CSV file with individual scores – no aggregate reports
- Will have RIT scores & will compare with MAP Growth – but extensive caveats for using this for any high stakes decision making
- If compare this to other methods, RIT variance is less than NSCAS distribution? Less variance but greater skew
  - RIT has adjusted ability? How much of adjustment is causing the pull-in?

- TAC comfortable with equipercentile approach. Regarding scores – individual scores better than providing district or building aggregate reports – look at individual student growth – will have achievement level information
- Messaging is important – schools believe 2021 data will not tell them much – large districts will make comparisons to spring MAP- is MAP RIT comparable to NSCAS RIT
- Other questions to consider from district perspective:
  - Will I still have goal setting worksheets?
  - Will I have learning profile reports at student, teacher, school, and building level?

Will have to explain why schools are not getting those pieces for 2021. Communication around limited utility but this is how you can use the data. Important for building future assessment system - getting field test with as much participation as possible for linking study.

MATH ALD Validity Study – efficacy of content developers' alignment of items to ALDs against empirical alignment of items based on item difficulty data & investigate the use of ALDs as the epicenter of the test score interpretation validity - results encouraging -will rerun the study after revise ALDs

- TAC
  - Like approach –
    - Seems like a bookmark method of standard setting – laying out items in order of difficulty & ask where students can no longer do work
    - How do you differentiate this from bookmark? Added benefit?
      - In bookmark trying to find out where is point just proficient student would be able to answer *minimum need to know* but this test ID matching procedure doesn't look at students but items – if good alignment information you cannot look at all the items, but if any tagged, use information to calculate cut scores where cut scores are optimized – putting interpretation & use front & center in process
    - Requires very detailed ALDs? Yes, very explicit so each standard has this articulation in terms of range of performance continuum – trying to get at progression of learning – what skills underlie standards, create ALDs that are more useful for instructional purposes & understanding standards
    - DOK & item difficulty is synonymous? Not a relationship between them - true when looking at fixed test form, but when explicitly building items at or above DOK levels, when purposeful manipulation will see DOK has influence on item difficulty in many grades/content areas. Trying to make sure that building ALDs can be validated to support item development and teachers can understand why.

            - Concerns about 2018 performance levels – telling stories difficulty when students earned 90+% on MAP Growth but not CCR on NSCAS – will performance levels be reset/renamed for TY?
                - Naming problem on top side – no intention saying “on track” didn’t mean on track for CCR
                - Will need to set new cutscores, but no answer on naming
            - CCR claim is associated with predictive CCR performance, not retroactive look at achievement claim? Benchmark is better rather than CCR (elementary grades) – names based on CCR standards rather than performance – reexamine names since setting new cutscores – intentions need to be addressed
            - Generally like methodology/approach – advantage is intentional design of assessments to support ALDs & achievement levels – matter of efficiency (bookmark created ALDs after cutscores set) with variations later used ALDs up front – this approach brings assessment, ALDs, & standard setting up front = cohesion & efficient with Round 1 work
                - From an efficiency standard, trying to make sure test has the capacity then can look at the empirical data and make sure what set is reasonable. Item feature characteristics can influence common vs uncommon features, conceptually easy things can scale out much more difficult for various reasons. Are additional things that can factor into the judgment, don’t know if want to purge those from the scale but naturally will be disconfirmation
            - Second element of disconfirmation is polytomous items. Achievement level descriptors match item with score point but needed panelist to break down how score point was achieved. Overall from efficiency standpoint, this approach puts you in better position when get to empirical data. Gives better chance on front end from design to make sure have item bank to support what have at the end.
            - When you take each grade level standard and articulate at each achievement level, the progress across grade levels. To extent articulate standard at lower level are you moving to prior grade. So look within each grade level, not testing prior grade level. Trying to keep test focused on grade level. Intersection of context plus how item presented

- Science – discussed PAD, task development workshop, SCILLSS, ToA, exploration of classroom impact of the workshop training
    - TAC:
        - Did see differences in student performance based no task? Not field-tested but will get feedback based on classroom use.
        - What about bias? Reviewed with bias & science content by outside experts
        - Curious to know beyond perception of teachers regarding student performance
        - Are there follow-up questions regarding teachers use, how they work with students – teachers attend PD, but does it really impact teaching?
            - Follow-up survey will ask how easy it was to use? Did it change the way of thinking about assessment and how impacted other teachers?
        - Interested in learning: if there were changes to curriculum & instruction, to what extent was student achievement impacted, changes in teacher attitude & behavior, students’ attitude/behavior in using this approach – greater engagement – what changes made to instruction as a result?
        - Goal seems to be to increase frequency of high-quality science tasks used in classroom – growing # of teachers creating growing # of tasks – another goal is getting this to all teachers; now you have a model, add to it – clearinghouse for teachers to use to add items
    - NDE: Tasks are like those they will find on the summative – allows them opportunity to experience types of testing/thinking vs discreet knowledge