Minutes – Nebraska Technical Advisory Committee meeting from September 20, 1019

**Location and Time:**

Hilton Garden Inn, Lincoln

September 20, 8:00-3:30

**Attendance**

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

**Approve Minutes March 22, 2019** – *Approved minutes as submitted*

**Science Measurement Model**

* NWEA presented information regarding a study that examined the appropriateness of different measurement models for the new NSCAS Science assessment based on the Spring 2019 pilot test data. Measuring whether students can do science rather than if they know discrete facts is goal of measurement model.
* A unidimensional IRT model most commonly used, but new practice for test companies is to use a multi-dimensional IRT, which needs a large sample and more time to estimate parameters.
* Original review was to look at how students answered questions but went down analysis path (not best sample for it). There were 2 types of dimensionality analysis: principle component analysis and cluster analysis and unidimensional and multidimensional IRT based models. The multidimensional IRT models had 2, 3, and 5 dimensions.
* The model fit statistics suggest that the 3-dimensional IRT models aligned to the DCI, SEP, and CCC dimensions provide the best fit to the pilot test data; however, to determine if the unidimensional IRT model is best fit overall, NWEA will conduct a simulation study after field test data has been collected. Results will be shared following the field test.
* Study used complex model to explain data which would be used by NDE to make final decision regarding the test.
* Small correlation explains why difficult to choose model. Correlation is moderate but not consistent across. There is no clear pattern.
	+ Big question is what are the correlations between items and what are the parameters actually doing.
	+ Operationally, unidimensional is easier with quicker results.
	+ We want to be true to conception of 3-D science learning in measurement model even if final score doesn’t communicate the complexity.
	+ Need to consider if model will be more accurate than others in school level accountability system.
	+ Hierarchical model may be better to use relative to simulation. It is a stronger path from general to simulation ability as a way to test robustness.
	+ Introduce a modified 2PL model. It’s more about utility – what will you do with the information? This is close enough we can live with it psychometrically but can it be communication to districts.
	+ Hold off on final model – see what other states are doing and modify it for Nebraska.
	+ From educational practitioner side, would want to know if this measures achievement, and what things a school could influence. The test is all about skills that can be taught. The test does address science standards and will follow the science standard blueprint.
	+ Regarding science assessment, NDE is working with a three-year plan with some standards assessed more often than others, but all standards assessed.
	+ Challenge is the number of combinations possible – is the model robust to allow sampling that addresses the standards.

**Science Field Test Approach**

* NWEA presented information regarding the Spring 2020 NSCAS Science Field test. The assessment is comprised of 11 Grade 5 and 10 Grade 8 scenario-based science tasks constructed from phenomena. Each task has 3 or 4 prompts. The field test is not adaptive, but instead has a fixed form. Following the field test, an embedded approach will be used to field test newer tasks starting in Spring 2021.
	+ Intentional stratification is used and stratifies across the pupil level. The engine is accurate at doing this but will be good to have a backup with linking items.
* NWEA prefers Option 2, in which forms are administered to equivalent groups of students using a stratified sampling approach. Linking sets are included in each form. This option depends on the available tasks and the ability of the engine to create equivalent groups using stratified sampling and to administer six (6) forms.
* The operational test will have 10 tasks. Later there will be 4 common tasks (overlapping) that will be used for equating purposes.
	+ Stratification will add complexity to hitting targets due to intentional engine sampling.
	+ If one set of items do not test very well or is a hard question, an item may be removed if it doesn’t destroy the continuity. It may result, however, in losing an entire set if it is a critical item.
	+ The state of Washington has model like this one. Look to their configuration/research.
	+ States field test more questions than needed but this is different. Option: look at multiple testlets for one phenomena with larger sets.
	+ Looking at embedded field testing, from researcher standpoint, placement is key. There should be variety to placement and a way to pause the assessment.
* Engine can place tasks in the same position an equal amount of time; however, if placed in different locations it might affect the operational performance.
	+ Field test with paired position; we can look at stratified assignment and see whether or not there are differences.
	+ If can randomize order of tasks, NWEA should try to do that.
* Concern about test security: items leaking out if testing is spread out over several weeks. While the phenomena may be memorable, the underlying tasks are not obvious.
	+ The design with linking tasks is stronger than a design that only depends on sufficient number of students.
	+ NWEA can evaluate item parameters during the first weeks of testing compared to the last week of testing to see if there is drift.
	+ After the field test is completed, NDE will look to see what level of attrition there is.
	+ Concern about this kind of test (i.e., phenomena/task based) causing a narrowing of curriculum. NWEA stated importance placed on standards/items rather than phenomenon will prevent this.

**Test Security**

* NDE annually updates a comprehensive plan for test security for the NSCAS system. Part of the security plan includes services from Caveon. Prevent, Detect, Deter, React. NDE asked how to communicate the services provided by Caveon and the best way to use data provided.
* TAC did not get a copy of the Caveon report from last year to review in advance.
* Presently there is no information shared regarding Caveon’s role in test security.
* There has been no evidence to cause NDE to launch an investigation based on data from Caveon. All investigations have been due to districts self-reporting.
* Caveon will develop a Security Incident Response Plan that helps the State when a breach happens with the possibility of sanctions.
	+ Share with districts that Caveon exists as part of the security plan. It services a purpose as prevention, protection, investigation, and enforcement. Emphasize the prevention aspect. Share the procedures in place, but do not have to go into detail. Let the districts know the procedures used are valid and are used in accountability.
* The security plan and using Caveon protects the integrity of the data to make sure it is a level playing field for all students.
	+ Release a lessons learned document based on what Caveon finds and reminders on how to administer tests.
	+ Target of communication needs to be at the level of greatest risk of malfeasance.
	+ Communicate with districts before begin to act on data from Caveon. NDE should create criteria in advance. Does the instance identified put school or district at significant advantage could be where investigation is called.
* Patterns over years are also important in determining when to investigate.
	+ When NDE investigates, work with media to inform about the policies, what NDE maintains, and how the data is interpreted within context and maintain sensitivity.
	+ Instances that meet the bar should be investigated and people made an example of so districts know NDE looks at test security.
* Caveon has a conservative threshold. Priority is data forensics list and will uncover why an anomaly happens. The strength of each piece varies.
	+ It is important that we have a trustworthy source of data.

**NSCAS Alternate Science**

* NDE will be conducting a NSCAS Alternate Science Field Test in Spring 2020 based on the Nebraska College and Career Ready Standards with Extended Indicators. The science alternate tests assesses students in grades 5, 8, and 11.
* All items are scored dichotomously with three (3)-answer questions.
* Federal requirements maintain NDE must use parallel language on reports. There is some latitude with communication but levels will not change (i.e., developing, on track, college and career ready).
	+ Discussion regarding three dimensions in science using task-based model clear this was difficult for students. Compromise was using stimulus with multiple questions. The proctor reads the prompt each time.
* Creating one test to serve 1% of the student population that is also more diverse than the rest of the student population is a challenge.
* NDE received a waiver to administer a standalone field test in the 2020 school year in order to construct operational assessments in 2021.
	+ A national group working on alternate assessment in the Enhanced Assessments Grant area. As move forward, see if they have anything out publicly to inform what NDE/DRC is doing in development stages. (EAG product called I-SMART)

**ESSA Business Rules Review**

* In an effort to focus resources to the schools most needing support and improvement and to ensure compliance with federal accountability requirements, the NDE designates schools for comprehensive support and improvement (CSI), targeted support and improvement (TSI), and additional targeted support and improvement (ATSI). These designations follow outlined policies written in the ESSA plan, and share the AQuESTT indicators used for classification (i.e., Needs Improvement, Good, Great, Excellent).
* CSI schools are lowest performing 5% of Title I districts. TSI are generally lowest 25% and the cut offs for ATSI will be made clear following these designations.
* For Stage 1 districts must have at least 10 students in the class who take ELA and Math for each indicator.
	+ NDE should use cohort approach over 3 years to help with volatility of data for schools with n < 10. Presently, for small schools, use District data when n < 10 students, but cannot use it some areas.
* Will have 6 reviewers for AQuESTT. If all reviewers do not agree on the rating, there must be agreement on 3 of the tenants.
	+ What was rate of convergence between ratings and how schools rated themselves? There was great variance among raters during first year of process.
	+ Updated process: schools are able to submit narratives to demonstrate why their rating should be increased. Have set of business rules and schools have an exemplar to make give guidance to schools.
	+ Goal is to winnow down number of schools; more qualitative data to identify schools in need of improvement. If the program is not helping to triage quickly, then perhaps need to lower threshold.
	+ Federal government said cannot change status calculation. Suggested start with lowest 25% then narrow to workable number with EBA or science results.
	+ If at 35-40% may want to reduce to lowest 1/3 to narrow pool.
* With accountability, not knowing where the target is, is a problem.
	+ Must have credibility in the field. If districts are saying they cannot influence AQuESTT, must look at this.
	+ NDE will insure target cut points are clear and return with more complete analysis of classification, movement and stability.

**AQuESTT Future Indicators**

* Three major changes have spurred enhancements to the AQuESTT system. First, the State Board’s 2016 Strategic Vision and Direction set ambitious, long-term goals aligned to the tenets of AQuESTT. Nearly simultaneously, the state embarked upon higher, more rigorous college and career ready standards and assessments. Finally, the passage of the Every Student Succeeds Act provided significantly more flexibility for states to innovate with their school accountability systems. Taken together, these three changes provide a window to achieve the broader vision of accountability, and to signal progress toward schools in meeting the State Board’s aspirational goals outlined in the Strategic Vision and Direction.
* Looked at several models included with growth scale but want usable and valid data that allows conversations regarding X number of students moving up.
* State board wants to be more comprehensive.
* NDE needs to verify that status takes into account proficiency.
	+ If cannot define it, is it meaningful. To what extent does a new indicator correlate with other indicators? Does it add value to the profile?
* Idea of using ACT and PreACT - what is the metric going to be? Will it be on a scale? Will it be student growth percentile?
* Concentrator versus Completer versus AP Course – Will there be unintended consequences (i.e., curriculum)?
	+ Would add certification to the list – IT, HR, CAN, etc. Can get these early
	+ To narrow down ANSI and ICE, have accreditation programs that have certifications that are accredited.
* Assessment tool used for college readiness – Do students take ACCUPLACER to validate On-Track ACT?
* Can educator effectiveness be measured through gap closures?
* For an indicator around discipline, do we consider expulsions and suspensions?
* What shall be shared in at the next meeting?
	+ Need indicators to see how can operationalize.
	+ Combinations to create a composite.
	+ Weighting opportunities and extent of the data.
	+ Back track it and see how stable it is.
	+ Fair doesn’t mean equal. What is fair for larger or smaller schools does not mean it is equally applied across all of them.
* At the next meeting, TAC will dig into the numbers.

**Next Meeting/Adjourn – there will be a doodle poll for late March or early April**