Meeting Notes
ISBE Technical Advisory Committee for Assessment and Accountability
June 4-5, 2024

Participants

ISBE	Tiffany Burnett, Rae Clementz
TAC	Jeff Broom, Ellen Forte, Laura Hamilton, Jim Pellegrino, Mike Russell, Diana Zaleski
Center	Chris Domaleski, Will Lorié, André A. Rupp
Pearson	Mary Allen*, Amanda Fitzgerald*, Eric Moyer*, Yong Luo*

** virtual participants on day 1 only*

Tuesday, June 4, 8:30-4:30 CDT

Welcome, March Meeting Recap, and Agenda for Current Meeting

After some brief introductions, Chris D. reminded everyone that several of the shared documents are confidential. TAC members should not share these materials outside of the TAC.

ISBE Update

Protest around ACT as the new high school assessment for grades 9-11 has been resolved; ISBE posted notice of award to ACT, which will be in place for six years. The state Superintendent recently announced that the Science portion of the ACT will be used for the science assessment work in grade 11.

Unified standard-setting work has commenced. As a reminder to the TAC, IL has a law that requires the state board to use part of a free college entrance exam for accountability purposes.

The first administration of the ACT will be spring 2025. The ACT is primarily a digital, online test with paper versions only offered as an accommodation. The TAC wondered whether peer review resubmission is required for science (yes), whether ACT Science will pass peer review alignment requirements (ACT Science passed peer review in Wisconsin), and whether the ACT Science section can be considered a meaningful science assessment in high school. They also noted that this also requires the modification of the existing contract (yes). ISBE sees value in signaling the importance of science education because the assessment is being given at grades 9, 10, and 11 with pre-ACT forms being administered in grades 9 (non-secure) and 10 (secure).

The TAC also wondered whether high school science education might shift to align more closely with how ACT assesses science, especially given that ACT science does not have much of a disciplinary content knowledge focus. This question also opened the door to ACT presenting their work at the TAC meeting. Rae is interested in the ACT team presenting to the TAC.

The Center suggested three issues should be addressed:

1. Alignment
2. Performance Standards
3. Accessibility

All these issues typically come up prominently in peer reviews. In terms of accessibility, ACT compared favorably relative to the College Board. ISBE wants to frame this as an opportunity.

In terms of graduation requirements, there are no minimum scores; rather, simply taking the ACT is sufficient.

The use of the ACT also opens the door to alternative computations for growth, specifically SGPs, given that there is an alternative system of assessments in place for high school. Rae signaled that the Superintendent is interested in considering growth in high school (computing and reporting an SGP was one of the requirements of the RFP).

Illinois Assessment of Readiness (IAR) and Illinois Science Assessment (ISA)

Yong Luo from Pearson provided a walkthrough of the different aspects of this work. The high-level points are not reiterated here.

Form Construction

Pearson reviewed the form construction process for the IAR and ISA, including test and calibration specifications and relevant operational constraints, and discussed psychometric targets for Test Information Functions (TIFs) and Test Characteristic Curves (TCCs). Note that the IAR is pre-equated while the ISA is post-equated.

One highlight of the presentation was that the TIF was essentially maximized around the “proficiency” cut on the scale. Yong discussed four core approaches to constructing the operational forms; the fourth option (the most general one) envelops the other three. The resulting question about which of the four approaches is best in the abstract cannot be decided upon but explored via simulations.

Yong noted that they inherited the scale properties from PARCC, resulting in scale forms that have a difficulty range of approximately -4 to +4.

<u>TAC Discussion</u>

The TAC appreciated the work but questioned whether the best empirical approach might not overlook critical issues of design to ensure that the blueprint is sufficiently covered. Content considerations are important in this context. In other words, psychometric criteria should not be the only ones to use, even if more sophisticated computational approaches are used.

The TAC affirmed the importance of adherence to the TCC and TIF, and having sufficient precision across the range of the scale.

TAC members also underscored the importance of understanding the different uses of the assessment information, which could be expressed via a general framework, for instance. This issue is further complicated by the item coverage of the ACT Science assessment.

There is also a risk of overcorrecting solutions across the years if decisions are made basically with empirical considerations only. The TAC supported Pearson's use of the prior year's form as the psychometric target, rather than an earlier baseline.

There is also a notable difference in reading load in the current ISA and the envisioned ACT Science test. However, these kinds of considerations can be investigated with response time analyses.

Procedures for Calibration and Equating

As a reminder, the ISA is post-equated and all modeling is done via a Rasch model using Winsteps. Internal analyses have shown very promising results with very few items displaying large amounts of DIF. Pearson reviewed the procedures and specifications for calibration and equating.

Part of the discussion focused on the concept of the displacement statistic, which is an estimate of the stability of the item difficulty parameter estimate under different estimation conditions.

TAC Discussion

The TAC wondered whether there is a qualitative review component to the displacement analysis to understand the root causes of these phenomena better. Pearson noted that, indeed, there is a content review step included in the work. The TAC also wondered whether this content review extends beyond the individual items and includes the content composition of the full, resulting blueprints. The TAC also noted that there is a risk sometimes in sequences of smaller adjustments, which can distort scales over time in the aggregate.

The consensus from the discussion from the perspective of the TAC was that the practices at Pearson align well with established professional practices. Moreover, Pearson seemed enthusiastic about the results they have already seen, since they have seen many examples of stable item parameters and equating results.

Differential Item Functioning (DIF)

Pearson reviewed when and how DIF analyses are produced using a flowchart, a summary of DIF findings by program, as well as a description of what happens when items are flagged for C-DIF and how results are documented.

TAC Discussion

The TAC wondered whether DIF analyses are run on operational data also or only on the field test data. Pearson noted that, unless there is some unusual concern or situation, no additional DIF analyses are done using operational data, given that field test samples are large and representative. Later, the TAC recommended including an additional cycle of content and bias review.

TAC members cautioned against keeping items with C-DIF and recommended either revising and re-field testing it or removing them. The TAC also wondered what the root causes of some of the DIF patterns were. Pearson noted that there are no systematic differences across the different DIF analyses.

Given the large number of comparisons for DIF analyses, there is also a higher chance for type-I error. The TAC suggested that it might be valuable to re-evaluate the DIF under operational use and, later on, possibly to over-sample C-DIF items.

TAC members also suggested more in-depth analyses using alternative variables that might get closer to the root causes of DIF (e.g., curriculum / textbook choice).

Similarly, DIF could also be seen as an indication that the model is too limited and the TAC suggested exploring alternative models. Pearson appreciated the notion but was hesitant to commit to doing this since a confirmatory analysis would be needed.

TAC members also recommended using the scale version of the DIF effect size rather than the coarse-grained A/B/C classification scheme. This is currently not done once the classification has been made but was seen as a valuable alternative by Pearson that is worth exploring in the future.

Exploring Student Choice on the IAR

Yong Luo from Pearson shared what they have learned from other states concerning student choice on assessments and identified the key constraints that should be considered for implementing student choice. He reviewed the benefits and tied these to the statistical issues of data becoming missing not at random, and scores not remaining comparable based on the professional literature. He later also described what modeling approaches are available in the literature to address these issues.

The TAC pushed back on this framing, though, given what the initiative around student choice is supposed to do.

The TAC also recommended defining use cases in question very clearly as the issues of choice play out very differently across, say, standardized assessments, performance tasks, and portfolio assessments. The TAC then discussed the issue of choice specifically in the context of constructed response items, including essay-type responses.

ISBE underscored that one key motivation for considering choice on the state assessments is to provide more equitable opportunities for all learners. Options that were discussed included hot vs. cold reads, topic selection, response options, and the like.

A discussion ensued about what the literature can tell us about the relationship between DIF statistics and understanding the impacts of student choice. The TAC expressed a desire to learn more about the relationship of student choice and student performance on various assessments.

The TAC also cautioned that there might be a risk of additional confusion for teachers if student choice mechanisms were implemented. TAC members advocated for scaffolding to help teachers and students understand the implications.

The TAC further advocated for careful consideration of the unintended negative consequences of any model that involves student choice. TAC members recommended some qualitative studies of schools or districts with different cultures around providing student choice in classrooms before students take standardized assessments with different degrees of choice.

A potential next step is to outline research studies / supplementary pilots tests to help address the challenges and opportunities of student choice.

Unified Assessment Standard-setting

Will Lorié from the Center provided a general update on the broader plans to implement the Unified Standard Setting process as discussed in March. He reviewed the rationale, the plans, timeline, and methodology, and then asked for input.

<u>TAC Discussion</u>

The TAC wondered what the definition of coherence in this context is. ISBE clarified that they want it to be vertically coherent and have face validity. ISBE noted that the tests are not primary content tests but, rather, skills-based tests.

Regarding the process, ISBE wants to know from the policy group how they are thinking about coherence. Drawing this out from the constituents during the workshops is, indeed, one of the goals of these workshops. Various aspects of coherence were briefly discussed, including coherence or consistency in wording, consistency in terms of specific expectations across grades, and so on.

The TAC wondered whether a particular method had been already determined (no). The process of going through a range of performance to determine what appropriate thresholds are is particularly challenging in general. ISBE noted that Pearson will be their partner for the standard-setting work.

The Center reiterated that the idea is to not just get guidance about the high-level policy performance level descriptors (PLDs) but go into the workshop with sufficient scaffolding to get sharpness around the PLDs. Part of this work will involve reifying, whenever possible, that these PLDs are about determining academic skills, not all kinds of 21st-century skills that, although important, cannot reliably or meaningfully be measured by an assessment.

In exploring the highest priority issues to address, ISBE noted that the Superintendent has said repeatedly that the rigor / the performance expectations are too high. This led to a discussion about the state's theory of action and what it says about the relationship between revised PLDs, changes in practice, and changes in performance. Later, the TAC noted that this comes back to the coherence of the ecosystem of curriculum, instruction, and assessment. A TAC member noted that even though, on the surface, the workshop goals seem to be about assessment, the work is not just about assessment (for it to be meaningfully implementable in the long run).

Related to this, the TAC wondered whether there is a concern that this effort may be interpreted as "lowering the bar" of performance (maybe). This, along with other aspects of the theory of action/change, can be mapped backwards explicitly through use cases that illustrate desired changes. ISBE noted that this work is tightly tethered with the accountability work. Right now, though, ISBE finds that they are sending a "loud, but not very differentiated signal."

The TAC also noted that there is a risk that people with different kinds of power relationships - in particular teachers and school or district leaders - may not be comfortable (and used to) contributing equally. This requires ongoing community building, both at the outset and at every possible opportunity. A later discussion looked at different sides of this issue.

The TAC noted that transparency about the design constraints that are already in place is critical - what can be changed and what is fixed. This could furthermore be interpreted with respect to the current state and with respect to the future. It would be important to describe how the PLDs can be used to design future assessments and, perhaps, reimagine the IL standards that have been in place for about 10 years. Related to this, showing a roadmap that illustrates the place that this workshop has in the longer-term strategy could be helpful.

The TAC applauded that ISBE is taking such a collaborative approach for developing the PLDs, which they noted is rare among states.

Representing Academic Achievement in the School Accountability System

André Rupp from the Center provided some options for calculating academic achievement based on earlier discussions, demonstrating the impact of alternative computational approaches using a sample of legacy data.

TAC Discussion

ISBE clarified that the cut scores for all grades and all subjects will change, that the timeline for the cut scores has no flex, but that the timeline for implementing changes in the accountability system is more flexible. ISBE expressed an interest in exploring a decision tree model for supporting decisions about supports.

The TAC suggested exploring how sensitive the system is to change, and putting to paper timelines and rationales for the changes to the program and accountability system. There was consensus on sunsetting the annual targets in 2025.

Wednesday, June 5, 8:30-11:30 CDT

Graduation Rate

ISBE has identified some challenges with how the graduation rate is currently included in school accountability. For example, graduation rates form a very narrow band, so it is difficult to differentiate schools on this metric alone. Additionally, students with disabilities may remain enrolled until age 22, which impacts the computation of the 4-year graduation rate.

ISBE described these and other core issues of concern while Chris Domaleski from the Center provided a brief review of how other states handle such issues in statewide school accountability settings.

TAC Discussion

A clarification question centered around what ‘meaningful differentiation’ means in this context. In essence, there is very little variability in graduation rate itself.

TAC members wanted to know whether the weight of 50% in IL could be adjusted. This is technically possible but would require more stakeholder engagement. Moreover, now that there is a possibility to model growth in high school via the ACT, this information could be used to shore up the evidence about development in high school.

ISBE noted that this connects to the performance of the accountability system at the high school level. In essence, different weights and compositions are needed at the elementary and high school level and a more diversified mix of indicators is needed at the high school level.

ISBE noted that there are real concerns that undesirable behaviors are incentivized (e.g., district leaders removing cluster programs that would move students out of the graduating cohort) through the narrow definition of what counts as a diploma, the high weight that it receives, and the fact that alternate high school diplomas do not exist in IL. This suggests that alternative accountability systems for alternate high schools would be beneficial - something that should be advocated for during ESSA reauthorization.

The TAC noted that there are alternatives for tweaking the system such as raising the ceiling or lowering the floor, using rolling averages across multiple years, and several other examples.

The TAC had a few questions around the methodological focus. This included a question about whether the schools with similar/identical graduation rates are indeed similar enough (i.e., appropriately clustered).

During the presentation of alternative examples for managing college and career readiness (CCR) evidence from different states, the TAC wondered whether it would be more of a burden to start a complex system up or to maintain it. ISBE noted in this regard that this would likely also lead to time- and resource-intensive audits, which might be challenging to do.
The TAC noted that data on several of the indicator components (e.g., IB/AP level, postsecondary credit) already exist, which could be leveraged for analyses. A few recommendations from the TAC in terms of enhancing the indicator space for high school included adding a “sophomore-on-track” indicator.

TAC Planning Session

Next TAC Meetings

September 17-18, 2024
January 22-23, 2025
June 17-18, 2025

Closing Comments

With respect to potentially revising accountability indicators, the goal should be to create desirable behaviors even though this may not result in more differentiation.

The topic of student choice is interesting and it would be valuable to discuss this further as it has a lot of promise.

It would be helpful to understand the larger arc of the context for decisions, including envisioned timelines, touchpoints with stakeholders.

The ACT Science assessment seems to be a problematic choice given the efforts that have been made over time from the perspective of the NRC framework and the ISA. This involved a lot of IL science educators. It's a priority to address ACT at future TAC meetings.

Regarding coherence of the PLDs, assessment-related conversations need to be tied to considerations about curriculum and instruction ("PLDs do not belong to assessments"). In general, specialists are often not on the same page about how the standards are operationalized through different areas of practice.