

A Fuller Record

The April 23–24, 2026 Michigan TAC Meeting

Damian W. Betebenner

Prepared for the September 2026 TAC meeting · 10 September 2026

A comprehensive account of the April 23–24, 2026 meeting of the Michigan OEAA/OEE Technical Advisory Committee, organised by issue rather than by clock. It is built from the official summary minutes treated as a rough transcript, supplemented by everything the minutes could not carry: findings held in chart images, eight hidden slides, two speaker notes, and a participation total that appears on no slide. It is offered as a deliberate contrast to the official minutes — not as a correction of them, but as an illustration of what a different process would produce from the same two days.

Table of contents

Executive summary	2
What the committee left on the table	3
What could not be recovered	3
Session 1 — The Smarter Balanced interim assessments	4
What was brought	4
The adoption picture	4
What OEAA cannot see	5
The concordance table	5
Adoption advice	6
WIDA, off agenda	6
Session 2 — The department update	6
Session 3 — Accountability growth targets	7
The problem	7
What was tried and failed	8
What was proposed instead	8
What the committee said	9
What blocks it	10
Session 4 — The Michigan Test for Teacher Certification	11
Problem one: meaningfully different forms	11

Problem two: constructed-response endorsements	13
Session 5 — Subtest performance indicators	14
What Michigan reports now	15
The committee’s position, stated before any results	15
The study	15
What the figures show	16
What the committee advised on method	17
What the committee advised on reporting	18
A note worth keeping	19
What this summary could not recover	19
Sources	19

i What this document is, and what it is not

The Michigan TAC does not yet record its meetings. What exists for April 23–24 is a nine-page set of summary minutes, 5,613 words, drafted by Jessica Fenby at MDE and finalised about three weeks after the meeting. Those minutes are the source for this document, treated as though they were a rough transcript.

This is therefore not an independent account. It cannot be. Where the minutes did not record something, it is not here either — and the places where that shows are marked.

What it adds is everything the minutes’ format could not carry: the substance of a 52-page simulation deck whose findings live in chart images, eight hidden slides, two speaker notes, a handout total that appears on no slide, and the connective tissue between five sessions that a sequential record cannot express.

It is longer than the minutes by a factor of about two. That is the point of the exercise, not an accident of it.

Executive summary

Five sessions across two days at The Graduate in East Lansing. Three of them — growth targets, licensure form security, and subscore reporting — were the same question wearing different clothes:

When a measurement system is asked to carry more weight than its precision can bear, what should the state do?

Michigan’s growth model is asked to identify high schools fairly using a three-year projection horizon that leaves grade 11 students one year of runway. The MTTC is asked to produce “meaningfully different” test forms for fields that cannot supply enough examinees to equate them. M-STEP subscores are asked to distinguish areas of strength and weakness on a test deliberately engineered to measure one thing.

The committee's answers rhymed across all three, and they were more consistent than any single session's minutes reveal:

1. **Name the inference before choosing the instrument.** Asked what the MTTC's scores support, OEE gave a procedural answer — a passing candidate can be certificated. The committee asked for a substantive one and did not get it. The same move recurs in the endorsement discussion and, less explicitly, in the growth session's argument about what an Adequate Growth Percentile is *for*.
2. **Prefer less information reported honestly to more information reported badly.** Explicitly on subscores — the "least bad information" framing — and implicitly on MTTC forms, where the committee preferred "fewer forms with less overlap than many forms with lots of overlap."
3. **Raise the bar rather than soften the report.** On growth targets, against OEAA's concern that raw numbers discourage schools, the committee's answer was that adequate growth should sit above the 65th percentile, not that the number should be hidden.

Two sessions were operational rather than consultative: Jennifer Paul's on interim assessment adoption, and Andrew Middlestead's department update.

What the committee left on the table

Thirty-seven discrete recommendations. Two were accepted outright — pursue the Nearest Grade Average method, and support Ji Zeng's subscore work through direct collaboration. One was a recommendation *against* a proposal: do not build constructed-response endorsement tests. The remaining thirty-four are open.

Two things were promised for the fall meeting: a sample subscore score report, and the WIN Work Readiness Assessments vendor attending with its psychometric team.

The full register, with status, is [in the corpus](#).

What could not be recovered

Who said what. The minutes attribute every question and every recommendation to "TAC." Five members with distinct areas of expertise were in the room — Mark Reckase, Jennifer Randall, Gregory Cizek, Damian Betebenner, Martha Thurlow — and the record renders them as a single voice. There is exactly one exception, in the final paragraph of the last session.

This is not a defect of the drafting. It is a property of the format, and it is the single strongest argument for capturing the next meeting differently. A committee's advice is worth more when you can see which expertise produced it, and worth less when a reader must take "TAC Feedback" on faith.

Session 1 – The Smarter Balanced interim assessments

Jennifer Paul (OEAA), April 23. The agenda called this “M-STEP/Smarter Crosswalk” and scheduled it after growth; the minutes make it Session 1.

What was brought

Three things: the state of voluntary adoption of the Smarter Balanced interim assessments, a draft concordance table meant to raise that adoption, and — on no agenda item — a WIDA operations update.

The deck was three slides and ninety-one words. It existed to frame one question, which is on its final slide:

What information would educators need in order to better use the scores from the interims?

The adoption picture

Roughly 14,000 students across 21 schools in about 10 districts. The programme has been rebranded from a “pilot” to a professional learning opportunity delivered with the Michigan Assessment Consortium — a deliberate move away from language implying provisional status. Recruitment has been webinars, presentations, and direct outreach, with incentives offered, and the pitch explicitly does not ask districts to replace what they already use. Forest Hills and North Branch Area Schools are named as proponents.

! The number that appears on no slide

The participation handout is a plain table: district, school, tests completed in 24-25, tests completed in 25-26, difference, percent change. It does not print a total.

Its total is **14,756 tests falling to 14,495** — down about 2% — in a year when Forest Hills brought **six entirely new schools** into the programme.

Two movements drive it. Ruth Fox Middle School fell 1,418 tests, about half its volume. Whiteford Middle School more than doubled, up 264 tests. Round Elementary fell about 26%.

New districts joining are being cancelled out by existing districts contracting. That is a retention problem. Everything in the recruitment strategy addresses awareness.

The minutes record the 14,000 figure and the outreach activity, both accurately. They do not total the handout, because totalling a handout is not what minutes do.

What OEAA cannot see

Asked how many educators are actually participating, the answer was that the data is not available at the user level. School-level counts exist; teacher-level counts do not.

So OEAA cannot distinguish broad shallow adoption from narrow deep adoption, and cannot tell whether Ruth Fox represents a district disengaging or a practice changing inside one. Those imply opposite responses. MAC's barrier study, underway at the time of the meeting, is the instrument Michigan has for finding out.

The concordance table

OEAA's instrument for raising adoption is a table mapping Smarter Balanced interim scale scores to M-STEP scale scores, by subject, grade 3 through 7, and performance level. Two draft layouts were circulated, differing only in framing text.

The committee treated it as a communication artifact rather than a psychometric one, and its feedback clustered into three concerns.

Say what the number means. The worked example states that a grade 3 student scoring 2112 on the Smarter Balanced interim ELA test "could score" 1203 on the M-STEP. Asked what *could* means, the answer was "student would most likely score, or something along those lines."

That exchange is four lines in the minutes and it is the most important thing in the session. A concordance table's entire job is to convey a probabilistic mapping to people who are not psychometricians. "Or something along those lines" is not a definition, and the document's own caveat — that the table "is not a perfect prediction" — quantifies nothing. This is a validity question arriving disguised as a copy-editing question, and it is why the committee suggested a prediction study.

Put the conclusion first. Move the key takeaway and performance level information to the top of the document. Both drafts lead with explanatory narrative.

Show the work. Districts will want the formula, not only the lookup table — and, extending that, a tool into which a district could enter a theta value directly and get a concordant score rather than a band.

Both drafts still carried an unfilled placeholder: *"Additional blueprint information can be found at XXXXXX."*

 The suggestion that would make the table unnecessary

The committee asked how much the Smarter Balanced interim reports themselves can be modified, and whether the corresponding M-STEP score could simply be inserted into them. The answer was that this is a question for Smarter Balanced.

It is the highest-leverage item in the session. Every other concordance recommendation improves a document that educators must find, read, and apply correctly. If the interim

report carried the concordant score, the number would appear where the teacher is already looking, and the document would not be needed.

It requires asking a question rather than building anything. There is no record of it having been asked.

Adoption advice

The committee's adoption suggestions pointed at money and governance rather than at teachers: target district financial staff and school boards, and tell districts already using the assessment that they need not wait for reimbursement as they would with other assessments.

OEAA floated a stronger version — universal use with per-student money attached for instructional support. It drew no recorded response.

WIDA, off agenda

Two operational items. The **WIDA Alternate Screener**, for students already identified as having significant cognitive disabilities, has been released and announced; OEAA has asked users to notify the department so statewide usage can be tracked. That is voluntary reporting, which makes the resulting count a floor rather than a census — a real limitation for a small population where every case matters.

2026 WIDA ACCESS testing used a 10-week window and will use it again in 2026-27, with the traditional 7-week window returning after that. A two-year change in testing conditions across the whole EL population; whether it affects comparability across the boundary is not addressed in the record.

Session 2 – The department update

Andrew Middlestead (OEAA), April 23. The only session with no accompanying artifact, and the only one that is briefing rather than consultation. The committee made no recommendations.

Spring 2026 testing has been status quo with one exception: the new WIN Work Readiness Assessments. WIN will attend the fall TAC meeting, and the record notes explicitly that this includes its psychometric team — which is the substantive part, since a vendor presentation without technical staff cannot answer the questions a TAC exists to ask.

The state budget is in process. Asked directly whether MDE's budget would be cut, the answer was that this happens every year, the process is early, and nothing is final.

Benchmark assessments are still reimbursed by MDE. Schools are currently required to report the data, but that requirement is not in this year's budget. A reporting mandate without

an appropriation is a durable compliance risk: data quality degrades when reporting is unfunded, and the resulting dataset carries a selection problem that is easy to forget by the time anyone analyses it.

Leadership. Michigan is in the initial period of new State Superintendent Dr. Glenn Malayko, who is gathering feedback from districts. OEAA has had an initial conversation with him about reporting.

That last sentence is one line in the minutes and deserves more weight than it gets. Three of the five sessions at this meeting turn on reporting decisions — subscore reporting has a spring decision deadline, the concordance table is a reporting artifact under active revision, and the growth proposal would change what schools see and require an announced break in trend. A new superintendent's reporting priorities could reopen any of them.

Session 3 – Accountability growth targets

Chad Bailey (OEAA) and Jessalyn Smith (DRC), April 23.

The problem

Michigan's accountability index is overwhelmingly identifying high schools, and OEAA does not believe that reflects where the instructional problem actually is.

The mechanism is arithmetic. An Adequate Growth Percentile is the growth a student must sustain to reach proficiency within a set timeframe, and Michigan's content-area timeframe caps at three years — a cap set on reliability grounds and Michigan legislative policy. A grade 11 student has one year left. The same proficiency gap therefore demands a far more aggressive growth target in high school than in elementary school, and ELA shows a larger average AGP gap than mathematics.

Current research holds that a school needs five to seven years to meet growth. Michigan allows three at most, and grade 11 one.

As put in the room: current high school identification "indicates a systemic district instructional problem, but the identification needs to go back to the entire district experience, K-12." Other states have wrestled with the same thing. The model is working as designed; the design is the problem.

i Context the minutes do not contain

The deck's speaker notes, on the slide listing state-level uses of growth data, carry this:

Educator Evaluations were repealed from state law. Districts may now conduct those however based solely on local policy and do not report them to the state. School grades were also repealed from state law.

Two of the high-stakes uses that historically justified the growth model’s precision requirements no longer exist in statute. That materially changes the stakes picture for any argument about how much precision the model must carry — and it appears nowhere in the minutes, because it was never said aloud.

What was tried and failed

DRC set out to build a model producing a grade 12 target from a grade 11 prior, examining the grade 9–11 sequence in mathematics and ELA, restricted to final SGPs. Predictors SGP 2023 and SGP 2024; outcome SGP 2025. Linear, polynomial, and quantile regression, with main and interaction effects.

None of it fit. In the presenters’ words, “none of this is great.” Conditioning on performance level had been tried earlier and was “not too promising.” Piecewise modelling was floated, as was working between scale scores and predicting a pseudo grade 12 score to run SGPs on.

This is recorded at length here because negative results are the first thing a summary discards and the first thing a successor needs. The failure was of model fit on SGP-to-SGP prediction, which is a plausible outcome — an SGP is already a residualised quantity, and predicting one residual from two others discards most of the signal that made the original projection possible. Extending to scale scores, which DRC named as the needed next step, would test that reading.

What was proposed instead

The **Nearest Grade Average** method, previously used in 2018-19 for the PSAT 8/9 and SAT transition. Rather than modelling each student, it computes an average AGP for a norm group defined by assessment, content area, grade, and prior-year achievement NTILE(50), borrowing from the nearest grade where a full model is estimable.

NTILE(50) is a deliberate precision/stability trade: fine enough to give many small buckets — more precision than the logistic regression method’s handful of large ones — and coarse enough that each bucket averages 1,000 or more R-package AGPs, which is what buys the stability.

Method	Precision unit
Quantile regression (R SGP package)	Individual student score history
Nearest Grade Average	Group score history — many small buckets
Logistic regression (MI-Access)	Group score history — a handful of large buckets

It lets every high school grade project three years out, the same horizon as elementary and middle. Four Growth Index scatterplots showed alignment between R-package and NGA

AGPs to be “pretty close,” and application to show meaningful impact on overall rating — with the same magnitude of impact on the EL Index using WIDA.

! OEAA’s own open question

OEAA has not determined whether that impact is desirable, and said it would need time to research it.

This is the hard question of the session and it went unanswered. NGA was proposed to fix over-identification of high schools. Whether it *fixes* the problem or merely *moves* it depends on where identifications land after the change, and on whether the newly identified schools are ones anyone believes should be. A method producing more reasonable targets and less defensible identifications is not obviously an improvement.

The deck asserts that EL impact will grow as Michigan reaches five years post-COVID. The mechanism is only in a speaker note:

By 2026-27 WIDA NGA AGPs would be applied to grades 8-12. However, we are choosing to limit to data post covid (i.e., 2021-22 and forward). Meaning as of 2024-25, the most we can project out is 3 years.

The EL projection horizon is currently constrained by data availability, not policy. As the post-COVID window lengthens toward the five-year EL maximum, the horizon extends and the impact grows. The assertion is in the minutes; the reason is not.

What the committee said

The committee endorsed the method — “encourage OEAA to approach this avenue” — and then argued with the premise underneath it.

On low-achieving schools: growth-to-standard “is going to be fairly difficult,” and for those schools “the norm actually tells you more than the growth.” The 60th percentile is already adequate there, “which is a nuance for how growth is reported.”

Asked how much room OEAA had to adjust, OEAA said it dislikes reporting raw numbers because they discourage — growth that cannot be achieved reads as failure.

The committee’s response was to raise the bar rather than soften the report:

Above 65 should be goal of adequate growth. Your job is to get these kids to have more than a year’s worth of learning in a year.

The reasoning is that the 50th percentile is, by construction, a year’s growth in a year; a student who is behind and grows at the median stays behind permanently. If the target is proficiency rather than maintenance, the threshold has to sit above the median.

A tension the record leaves unresolved

Raising the adequacy threshold makes targets *more* aggressive, which is the opposite of what the session set out to do.

The committee's position appears to be that the timeframe should extend — via NGA — while the per-year expectation rises: more years at a higher rate, rather than fewer years at an impossible one. That reading is consistent with everything said, but the reconciliation is nowhere stated explicitly, and it should be settled before OEAA acts on either half.

Underneath sat a disagreement about purpose. OEAA's account is that the AGP “was an attempt to communicate the urgency to make improvements.” The committee would “rather see a system that supports all kids,” and held that “schools should not be focusing on proficiency; their goal is to produce and maximize growth.”

A metric designed to signal urgency and a metric designed to guide instruction are different instruments, and Michigan has been using one for both. Urgency-signalling tolerates unmeetable targets — that is what makes them urgent. Instructional guidance does not.

The committee also asked for two pieces of research nobody has: whether AGPs are actually used and used well, and whether elementary and high schools interpret the same data the same way. The second matters because the same AGP means something structurally different at grade 4 and grade 11, and if schools do not understand that, the over-identification problem has a communication component that NGA would not touch.

Finally, it suggested taking the work to MERA for a second, differently-situated review.

What blocks it

The AGP as currently defined is written into Michigan's approved ESSA plan. Changing it requires a plan amendment — “a large department-wide plan,” not something OEAA can execute alone.

OEAA believes the NGA change is a **technical update** within its own authority rather than a change to the AGP definition. That distinction is what makes the proposal actionable at all, and the corpus contains no record of it having been tested with federal programmes staff or with the U.S. Department of Education.

Because the change touches all grades, it would also require an **announced break in trend**. Asked how OEAA would explain a standard change and a trend break, the honest answer was that OEAA “hasn't gotten that far in planning” and is still deciding whether a new growth calculation should be done at all.

An announced break has to be announced before the results, not explained after them, and schools under accountability pressure will read any discontinuity in whichever direction suits them. It is currently unplanned.

Session 4 – The Michigan Test for Teacher Certification

Bridgett McDowell (OEE), with Nancy Hahn and Heather Klesch (Evaluation Systems–Pearson) and Jackie Thompson (Pearson Professional Assessments), April 23. The meeting's longest session — scheduled 1:00 to 4:00 with a break. No deck survives; this account rests on the minutes alone.

Two problems under one contract.

Problem one: meaningfully different forms

The new Pearson contract requires OEE to maximise the number of *meaningfully different* test forms for each field, and the programme must be inclusive, valid, reliable, secure, and legally defensible. The goal is a minimum of three forms by contract end.

OEE's working definition set a single high ceiling on overall item overlap between forms, with a slightly lower ceiling on the scorable set, applied uniformly across every field. Rationale offered was standard practice in educator licensure testing, programme integrity, and legal defensibility.

The committee rejected it, saying the ceiling was far too high to constitute meaningful difference — and then putting it in terms of the contract's own horizon: a handful of forms produced over several years, each sharing most of its items with the others, cannot reasonably be called meaningfully different.

The objection is not aesthetic. A candidate retesting inside the form repeat interval, against a small number of live forms that share the large majority of their items, faces very nearly the same test — which is exactly the security and fairness risk the multiple-form requirement exists to remove.

A note on redaction

The specific overlap ceilings, the form rotation parameters, and the per-group form counts are test-security sensitive and are not published. They are held outside this repository; the argument in this section does not depend on them.

The constraint that produced the definition

Low-incidence fields do not generate examinees fast enough. Statistical equating needs a minimum of 100 examinees per form. Re-using a single form raises security and fairness risk. Only a small number of forms may be live in any one testing window, and a candidate may repeat after a fixed interval.

High overlap is how you get more nominal forms out of a small item pool. The committee's position is that this is the wrong trade.

OEE's response had been to sort the certification tests into four groups by incidence. The two higher-volume groups build a new form once a field reaches the 100-examinee threshold, differing in which equating method applies. The two lower-volume groups do not equate at all and are standard-setting driven, with Pearson conducting the standard settings and item-level judgements used to build the operational forms.

Asked why the highest-volume group was permitted the form range it was, the answer was that this was what the old contract said — not that it followed from Michigan's own volumes.

What the committee designed instead

The core move: a single overlap rule across four groups differing by an order of magnitude in volume is the wrong instrument. Higher-incidence fields accumulate examinees faster, can support more distinct items, and should carry a *lower* overlap ceiling. Low-incidence fields need the tolerance.

Around that:

- **Roughly halve the overlap ceiling**, “in association with the other variables” — overlap is not independently settable; it trades against form count, which trades against volume and the 100-examinee equating threshold.
- **Fewer forms with less overlap beats many forms with lots of overlap.** The governing principle. The contract says *maximise the number of meaningfully different forms*; OEE had been reading that as maximising the count, and the committee is saying the adjective binds harder than the noun.
- **Start with a standard-setting form** — one scorable set and one non-scorable set of 20 items. Once a form has enough candidates, begin the next; non-scorable items that perform become scorable. This is the elegant part: every administration produces scores *and* field-tests twenty items, so the pool grows as a by-product of operating the programme rather than requiring a separate field test that low-incidence fields cannot fill.
- **Stick with 100 examinees, not 200.** Waiting buys precision on each form at the cost of halving the rate at which forms enter service, and form scarcity is the problem being solved.
- **If 100 cannot be reached in two years, escalate to the next group's process.** With the reasoning attached: educator preparation programmes will not change enough in under two years to move the numbers, but after two years the absence of change is itself information. Without such a rule, a low-incidence field waits indefinitely for a form it will never equate while re-using a single form — the exact risk the requirement exists to prevent.
- **Automated item generation** to enlarge the pool through cousin items — with the committee's own caveat that cousins are cousins, and that the focus therefore falls on the non-overlapping portion.
- **Define incidence categories for Michigan** rather than inheriting them. OEE is already examining annual volumes; the boundaries should follow from that distribution.

Two questions went unanswered. Pearson asked what definition and item overlap an equating methodology would require. OEE asked what annual volume yields 100 examinees per year. Pearson also asked whether equating data can be too old — answered “yes, but it is dependent upon so many variables,” with the resolution deferred to the next contract. For

low-incidence fields this may mean there is no volume at which equating is both feasible and valid.

The validity question underneath

! The most consequential exchange of the session

Early on, the committee asked what inferences OEE can defend for each of these tests. The answer: passing at the minimum cut score means the state can certificate the candidate, which allows a district to interview them.

That describes what the score *permits*, not what it *means*. It is a procedural claim, and it cannot be defended the way a validity claim is defended.

The committee asked for a substantive statement instead — what can you infer when a candidate does not reach the cut score, what does performance mean, what do MTTC scores represent?

Everything else in the session depends on the answer. “Meaningfully different” has no criterion until the inference is named: two forms are meaningfully different if they support the same inference equally well, and an undefined inference cannot adjudicate one overlap ceiling against another.

Problem two: constructed-response endorsements

OEE proposed constructed-response test forms for optional disciplinary endorsements — five in science (biology, chemistry, Earth and space science, physics, physical science) and four in social studies (civics, economics, geography, history). Candidates must pass the base science or social studies test first.

Nine additional tests, each requiring item development, rubric development, standard setting, and human scoring capacity, for fields whose volumes are by construction smaller than base tests that already struggle to reach 100 examinees.

The committee advised against it:

This creates a lot of work, will be hard to do, and will be expensive; suggest that this isn't necessary.

Pearson's question of how many CRIs per test went unanswered, because the premise was rejected.

Four alternatives were offered:

Transcript review — verify the candidate took coursework in the area. The cheapest option, and the committee named its weakness in the same breath: higher education institutions design their preparation programmes as they wish and are not aligned across programmes, so a transcript line means different things at different institutions. That non-alignment is

also part of why the endorsement test exists, so this trades a known problem for the one the test was built to solve.

Higher conjunctive cut scores on the existing test — require a minimum on the relevant sub-area as well as overall. Probably the strongest option: it converts an existing instrument into an endorsement decision at essentially zero development cost.

An obstacle nobody raised

A conjunctive sub-area cut requires the sub-area score to be reliable enough to bear a pass/fail decision for an individual candidate.

That is precisely the property the *next* session spent a morning establishing that M-STEP claim-level subscores do not have. The MTTC is a different instrument with different sub-area lengths, so the objection does not transfer automatically — but it needs answering before this route is taken, and the record contains no MTTC sub-area reliability evidence.

The two sessions ran on the same day and the connection was not made.

A randomly selected standard as an open-ended prompt, scored against a rubric. The most developed alternative. Social studies standards are not all written as questions, so the task would evaluate the skill of developing an argument — a skill applying across all four social studies fields, which is what lets one instrument serve four endorsements. Pearson already holds rubrics for other CRI items and the cut would be set from the rubric. Scoring capacity is the obstacle. A variant — performance items such as developing a lesson plan — was raised and not pursued.

An interview guide for districts. The most radical option, in that it declines to solve the problem with an assessment. OEE's own account is that a passing MTTC score "allows a district to interview a candidate"; the hiring decision was always the district's. This strengthens that decision without building an instrument the state must defend. It does not produce an endorsement, so it supplements rather than substitutes.

And underneath, the same move as before: **define the purpose**. What does having the endorsement on a certificate signify? The answer given — that the holder knows the content in the standards — is a low bar, and one that a conjunctive cut on an existing test could clear at a fraction of the cost of nine new instruments.

Session 5 – Subtest performance indicators

Ji Zeng (OEAA), April 24. The full morning of day two. The minutes date this session 04/204/26.

This session produced the most technically substantial work of the meeting and is the one the current minutes format handles worst. Roughly thirty of the deck's fifty-two pages are chart images carrying no text beyond a repeated title. Text extraction yields about 680 words; the findings are in the pictures. What follows was recovered by rendering every page and reading it.

What Michigan reports now

Subscores for ELA, mathematics, science, and social studies, in the Dynamic Score Reporting Site and on Parent Reports. ELA and mathematics report *claims*; science reports *domains*. Both use performance indicators — Above Standard, At/Near Standard, Below Standard — relative to the overall content cut. Mathematics Claims 2 and 4 are combined for reporting because of content similarity, so four claims yield three reported scores. Social studies reports raw points earned over points possible.

A known defect: for some subscores the At/Near Standard band cannot be drawn at all, because there is only one data point.

The committee's position, stated before any results

This determined everything after it. The argument arrived early and did not move:

Subscore reporting is used to identify areas of weakness and strength.

You have to have some difference between the subtests in order to use them to focus on, but the subtests are too similar, and it does not provide accurate information for focus due to similarity.

We do everything we can to make everything unidimensional, but people are demanding more information while wanting the test to be shorter.

Seems detrimental to provide that information from something you developed to not do that.

OEAA's own finding agreed: subscore reporting “doesn’t add any unique information.”

The committee then framed a choice — “we can either resist allowing them more information, or determine what method can we use that will give the *least bad* information” — and took the second option. Not “stop reporting subscores,” but “if you must, report the least bad thing, and be honest about how bad it is.” The operative phrase was a “happy medium between no information and inaccurate information from scale scores.”

The study

Zeng compared a Classification Tree with Proportional Odds (CT-PO) against the Smarter Balanced subscore indicator method, with K-modes as a reference.

True theta values from a Gaussian mixture model over 110 replications; operational item parameters producing 0/1 responses; operational raw-to-scale conversion applied; CT-PO trained on the first 10 replications and applied to the last 100 for classification accuracy; a

second set of responses over 100 replications for consistency.

Unidimensionality was checked by the Wainer et al. (2001) procedure — Cronbach's α per subtest, observed variance-covariance matrix, subtract the diagonal error matrix, eigendecompose the result and look for a single dominant eigenvalue.

Batch 1 varied component weights, mean vectors, SD vectors, and inter-subtest correlations of 0.65, 0.70, and 0.75, calibrated to science grades 5, 8, and 11 — 18 GMMs per grade, 54 fitted. Batch 2 widened the mean separation and SD vectors because Batch 1's conditions proved too alike.

What the figures show

 Recovered from the chart images

The empirical warrant (pp. 5–7). Three panels per grade overlaying an empirical kernel density estimate, a three-component Gaussian mixture, and a single Normal. At grade 5 the fitted Normal is $\mu = 1492.8$, $\sigma = 23.5$; at grade 11, $\mu = 1792.9$, $\sigma \approx 22.6$ –24.0.

The empirical densities are visibly multimodal and left-shouldered. The single Normal over-peaks the centre and under-fits both shoulders; the 3-component mixture tracks the empirical curve closely. **This is the evidence for abandoning the multivariate normal assumption** that Zeng criticises in the existing literature, citing Sinharay (2014) and Gorney and Sinharay (2025).

Unidimensionality (p. 18). Pooled across all conditions, the first eigenvalue is centred near **12.7** and spans roughly 10–16. The second and third both sit between **1 and 2.5**. A dominant eigenvalue by a factor of six to eight.

This is the committee's objection stated quantitatively. The data really is one dimension.

Within a single condition (pp. 19–20). Accuracy in quadratic weighted kappa clusters tightly at 0.63–0.66 for both methods, essentially on the identity line. A dumbbell plot over 25 samples shows CT-PO ahead of SBAC in every sample for Earth Science and Life Science, near-overlapping for Physical Science — a consistent win, but a small one, on the order of 0.01–0.05.

Across Batch 1 conditions (pp. 23–38). CT-PO accuracy spans about 0.56–0.78 against SBAC's 0.50–0.68, and **every point lies above the identity line**. Consistency shows the same pattern in a tighter band. No reversal at any tested correlation.

The deviance comparison against real data is the one table with a text layer:

Grade	K-modes	CT-PO	SBAC
5	71,894.50	71,675.26	97,032.84
8	73,622.17	71,757.89	85,159.56
11	61,396.68	57,574.52	77,622.43

! The caveat that matters most — and that only the figures show

Batch 2 varied the separation between the three latent component means:

Mean vector	Approximate accuracy (QWK), both methods
$\mu = [-0.25, 0, 0.25]$	0.10 – 0.35
$\mu = [-0.50, 0, 0.50]$	0.25 – 0.55
$\mu = [-0.75, 0, 0.75]$	0.35 – 0.70
$\mu = [-1.00, 0, 1.00]$	0.45 – 0.82

CT-PO's advantage is a function of how separable the three latent dimensions actually are. At the smallest separation the points hug the identity line and both methods produce classification far too poor to report on an individual student. As separation grows, both improve and CT-PO's margin widens.

Since the test is deliberately engineered toward unidimensionality — and the eigenvalue check confirms it succeeded — the realistic operating region is the *top* of that table, where the better method is still not good enough.

The committee saw this in the room. The minutes record it as one line: “when the means are close and the standard deviations are large, then both methods perform poorly.” That sentence is accurate and it does not convey the magnitude. The figures do.

What the committee advised on method

Model the joint density. The most technically substantive recommendation of the meeting:

Model the joint density of the three dimensions, and understand the density of them, determine what are the correlations between dimensions. Currently assuming linear correlation... Go from joint density to the marginals to placing the cuts in the marginals; you would want them to be faithful to the joint density.

The simulation parameterises the relationship among subtests as a single correlation value, which presumes linear dependence identical across the range. The empirical distributions are visibly non-normal, and a correlation coefficient cannot describe dependence in that setting — two variables can share $\rho = 0.7$ and behave entirely differently in the tails, which is exactly where classification cuts fall.

Establish scale, then cuts, then categorise from theta. An ordering discipline: categories should follow from a defined scale and defined cut points, not be produced directly by a classifier.

Review the item parameters before a third batch. The simulation draws theta from a mixture model but uses *operational* item parameters, which differ by subtest. Subtest-level

differences in the results may therefore be an artefact of the item pools rather than of the methods. The committee's related question — "why is life science different than the others; are other subtests more related to the total score?" — sharpens the point.

A further question went unanswered: Batch 1 imposes the same correlation on all three subtest pairs, which real data would not produce. If the true structure is asymmetric, the uniform CT-PO advantage would need re-testing.

What the committee advised on reporting

Here the committee reframed the problem, and this is the most useful thing it did all meeting. OEAA had been asking which method is more accurate. The answer was that accuracy is not the binding constraint — what a parent understands is, and nobody has measured that.

- **Report a level of certainty with any category, with a measure of error.** This is what makes "least bad information" defensible rather than merely pragmatic. A category shown without error looks like a fact; given classification accuracy in the ranges above, the uncertainty is not a footnote, it is the finding.
- **Run parent think-alouds and focus groups**, asking "What does this report tell you about this student?" and "What can you say about the student based upon this report?" The first elicits what a reader believes the report says; the second, what they believe they can conclude. The gap between them is where over-interpretation lives. OEAA is already developing a survey; the think-alouds were recommended in addition, because a survey measures reactions to a report someone has already made sense of, while a think-aloud catches the sense-making.
- **Offer drill-down reporting** — basic information, then subscore, then standard errors, then learning profile. A structural answer to the tension: the default view carries what a general reader can act on, and the uncertainty does not have to compete for space with the headline.
- **Bring a sample score report to the next meeting.**

There is no hard timeline for this work, but a decision must be made in the spring for the following year's results.

The one place an individual is named

JI Zengs's work is very sophisticated, and TAC would like to support her in her work. Damion Betebenner suggested that she work with him on this.

Both names are misspelled in the source. This is the only sentence in 5,613 words of minutes where a TAC member is recorded taking an action, and it is the mechanism by which the joint-density recommendation becomes something other than aspirational.

A note worth keeping

Among the reporting ideas recorded is one that reads differently five months on:

Could use generative AI model to input current report and then parameters around what you are trying to do or see.

The committee raised it in April 2026, unprompted, as a score-report design tool. It is the only mention of AI in the meeting, and it arrived as a suggestion about *documents* rather than about process — which is roughly where the argument in the companion brief begins.

What this summary could not recover

Three things, stated plainly so that no reader mistakes their absence for absence from the meeting.

Who said what. The minutes attribute to bodies. Five committee members were present on April 23 and four on April 24, with distinct areas of expertise, and the record renders them as a single voice. Every “the committee advised” in this document is faithful to the source and less informative than the meeting was. Nothing here infers an individual attribution, because a plausible guess would be a fabricated record — and a fabricated attribution in a technical advisory record is worse than none.

Tone, hesitation, and dissent. A minute records what was said, not how firmly. Whether the committee was unanimous on raising the adequacy threshold above the 65th percentile, or whether one member argued it against reservations from others, is unrecoverable. The record’s flat surface makes every recommendation look equally settled. They were not.

What was discussed and not recorded. A nine-page summary of a two-day meeting is compression of roughly ten to one. The MTTC session ran three hours and produced about 1,500 words of minutes. Whatever else was said in those three hours is gone.

None of this is a criticism of the minutes or of the person who wrote them. It is what happens when a two-day technical conversation must become a nine-page document three weeks later, drafted by one person, with no mechanism for representing a figure and no convention for attribution.

The remedy is not a longer document or a more diligent author. It is a different process — which is the subject of the companion brief, [Toward an AI-Native TAC](#).

Sources

All source documents are in the repository under `raw/2026-04-23/`. Structured summaries of each, with provenance and known defects, are in the corpus:

- [TAC Meeting Notes, April 2026](#) — the official minutes, and a catalogue of what they lose
- [April 2026 TAC Agenda](#)
- [CT-PO vs SBAC Simulation Results Deck](#) — the figure findings recovered here
- [TAC Growth Update Deck](#) — hidden slides and speaker notes
- [Interim and Smarter Concordance Table Examples](#)
- [Smarter Balanced Interim Participation Trend Data](#)
- [April 2026 TAC Interim Deck](#)

The complete recommendation register, with status: [Open Recommendations from April 2026](#).

Cited in the source materials: Wainer et al. (2001); Sinharay (2014); Gorney and Sinharay (2025).