

A Copula Approach to the Subtest Indicator Problem

Damian Betebenner

A response to [Session 5 — Simulation Studies on Subtest Indicators](#), written for [Ji Zeng](#) ahead of the fall meeting.

How this was written

I sat through this session in April and left with a specific, uncomfortable feeling: the work was clearly sophisticated, I could follow each analysis individually, and I could not have told you what question the whole thing was answering. That is not a criticism of the presentation. It is what happens when a two-day meeting gives you ninety minutes on a fifty-two-page deck.

Five months later, I did not re-read the deck. I asked the corpus:

- What did the deck actually establish? [CT-PO vs SBAC Simulation Results Deck](#), including the figure findings that reach no other record
- What did the committee advise, and what is still open? [Model the joint density of the three dimensions, Would varying the inter-subtest correlations change the result?](#)
- What is the decision OEAA actually faces, and when? [M-STEP Subscore Reporting](#)

That took a few minutes and produced something I could not have reconstructed from the PDF — in particular the Batch 2 gradient, which is the crux and which lives entirely in chart images.

This page is the result. It is the artifact the current TAC process has nowhere to put: a member's worked response, developed between meetings, addressed to a presenter by name. See `documents/ai-native-tac.qmd` for why that matters.

What I could not answer in the room

What is the research question?

The deck does not state one, and I think at least four are tangled together. Separating them is most of the work:

	Question	Status after April
Q1	Are M-STEP science subtest distributions well approximated by a normal?	Answered: no. Visibly multimodal; a three-component mixture tracks the empirical density where a single normal over-peaks the centre and under-fits both shoulders
Q2	Does CT-PO classify more accurately and consistently than the SBAC method?	Answered: yes , in every simulated condition, and with lower deviance against real data at all three grades
Q3	Is <i>any</i> method accurate enough to justify reporting subtest indicators for an individual student?	Not answered — and this is the one OEAA needs
Q4	What should the score report say?	Redirected by the committee to parent comprehension: Run parent think-alouds and focus groups on the score report

The deck spends most of its length on **Q2** and answers it convincingly. But Q2 is a horse race between two methods, and **Q3 asks whether the race matters**. Q1 is presented as motivation, but it is really evidence bearing on Q3.

The Batch 2 results answer Q3 and the deck does not frame them that way. As the separation between the three latent component means narrows, both methods degrade together — down to quadratic weighted kappa in the 0.10–0.35 range at the smallest separation, where the points sit essentially on the identity line. Since the test is engineered toward unidimensionality, and the eigenvalue check confirms it succeeded (first eigenvalue near 12.7 against second and third near 1–2.5), **the realistic operating region is the end of the range where the better method is still not good enough**.

That is the headline. CT-PO wins, and winning may not be sufficient. I would put it on the first slide rather than the forty-eighth.

Where a copula comes in

The committee's April recommendation was:

Model the joint density of the three dimensions, and understand the density of them... 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.

That is a description of a copula in everything but name. Sklar’s theorem says any trivariate joint distribution factors as

$$F(x_1, x_2, x_3) = C(F_1(x_1), F_2(x_2), F_3(x_3))$$

— the marginals F_i on one side, all of the dependence in the copula C on the other. Going “from joint density to the marginals to placing the cuts in the marginals, faithful to the joint” is exactly what that factorisation buys you. [Model the joint density of the three dimensions](#) is a copula recommendation that had not yet found its vocabulary.

1. It removes an assumption currently doing unexamined work

Batch 1 parameterises dependence with a single correlation — 0.65, 0.70, or 0.75 — applied across all three subtest pairs. Imposing a correlation matrix on latent normals **is** a Gaussian copula. It is a modelling choice, not a neutral default, and it is invisible because it never gets named.

In work with Henry Braun I selected copula families across 966 conditions spanning four longitudinal assessment datasets, by AIC with Cramér–von Mises goodness-of-fit and parametric bootstrap. The result:

Family	AIC-best
t	63.6% (614 conditions)
Frank	30.7% (297)
Gumbel	3.6% (35)
Gaussian	2.1% (20)

The Gaussian copula — the one implicitly assumed here — was the best-fitting family in about one condition in fifty.

The difference between Gaussian and t is almost entirely **tail dependence**: the tendency of extreme values to co-occur across dimensions. A Gaussian copula has none, asymptotically. The t-copula does, symmetrically, governed by its degrees of freedom. And classification cuts live in the tails. A dependence model with no tail dependence, used to study how accurately students near a cut get classified, is assuming away the behaviour that determines the answer.

I would not assert that this changes Zeng’s conclusion. I would assert that nobody currently knows whether it does, and that it is cheap to find out.

2. It separates two things Batch 2 currently confounds

Batch 2 varies mean vectors *and* SD vectors together. Both change the marginals, and changing the marginals changes the effective dependence structure at the same time. So the Batch 2 gradient — the most important result in the deck — mixes two causes.

Under a copula formulation they come apart cleanly. Hold the copula fixed and vary the marginals: that isolates what mean separation alone does. Hold the marginals fixed and vary the copula: that isolates dependence. The gradient becomes interpretable rather than merely visible.

3. It converts an open question into a parameter

The committee asked, and the record marks unanswered ([Would varying the inter-subtest correlations change the result?](#)):

Correlation between subtests are all the same because real data wouldn't show this... would it make a difference if you didn't have the same correlation between the three?

Under a single correlation parameter this is hard to explore. Under a copula it is a specification choice — a t-copula with a full correlation matrix and shared degrees of freedom, or a vine if the pairwise structure is genuinely asymmetric. The question stops being a limitation and becomes a condition to vary.

The related question — *why is life science different from the others; are other subtests more related to the total score?* — is answered directly by the per-pair Kendall's τ that falls out of the fit.

4. The data to estimate it already exists

This is the part that makes it practical rather than aspirational. All three subtest scores are observed **on the same students** at grades 5, 8, and 11. That is a directly observed trivariate sample. The dependence structure does not have to be assumed; it can be estimated.

And the marginals are already in hand — they are the kernel density estimates on pages 5–7 of the deck. Sklar's theorem says to take those as given, transform to pseudo-observations by empirical rank, and fit the copula on the ranks. The marginal work is done.

Where a copula will not help

I want to be straight about this, because the case is weaker than I would like in two places.

It may not move the numbers. In the companion sensitivity work I found that Student Growth Percentile classification was strikingly robust to copula choice — parameters stable, variants highly correlated, classification largely unchanged across specifications including deliberately mis-specified ones. There is a real chance the same holds here and that switching from Gaussian to t moves QWK in the third decimal place.

That would still be worth knowing. It would establish that the accuracy ceiling is set by unidimensionality rather than by the dependence model, which converts a modelling debate into a settled fact and points everyone at the reporting question instead. But it means this should be scoped as a bounded check, not a research programme.

It does not rescue Q3. If the three latent dimensions are not separable, no dependence model makes individual classification accurate. A copula changes how honestly the uncertainty can be characterised. It does not manufacture signal that is not there.

A caution on goodness of fit. In the 966-condition study, every parametric family showed statistically significant deviation at large n . With roughly a hundred thousand students per grade, a formal GOF test will reject everything. Family choice has to rest on AIC/BIC and on fit in the tail regions where the cuts fall — not on p-values.

What I would actually suggest

A bounded piece of work on data already in hand. Perhaps a week.

1. **Take grade 11**, where the CT-PO deviance advantage was largest (57,574.52 against SBAC's 77,622.43).
2. **Form pseudo-observations** from the three subtest scores by empirical rank. No distributional assumption; reuses the marginals already estimated.
3. **Fit Gaussian, t, and Frank trivariately.** Compare by AIC and BIC, and — more importantly — by fit in the corner regions where the classification cuts fall. Report per-pair Kendall's τ and the tail dependence coefficients λ_L, λ_U .
4. **Re-run the CT-PO vs SBAC comparison** with theta drawn from the best-fit copula at the *empirical* dependence, rather than from a correlation matrix at 0.65/0.70/0.75.
5. **Compare QWK against Batch 1.**

Two possible outcomes, both useful. If the numbers barely move, the dependence model is not what is limiting classification and the reporting question is the whole game. If they move, the existing comparison was conducted at a dependence structure the data does not exhibit, and Batch 3 should be specified differently.

Step 3 alone is worth doing regardless. Michigan does not currently know the tail dependence structure of its own science subtests, and it bears on every subscore decision that follows.

What that looks like in code

Against the `copula` package, in the idiom of the sensitivity study. It is a starting point to argue with rather than a specification — but it does run: everything below executes end to end on `copula` 1.1.7 with a simulated stand-in for the extract, so the API calls and the shapes are right even though the numbers are not yours.

Two things it needs from you: the extract itself as `dt_g11`, and `qgmm()` — the inverse CDF of your fitted mixture, which is marginal machinery already in the pipeline.

```
library(copula)
library(data.table)
```

```
## 1. The observed trivariate sample -----
## One row per student; the three science subtest scores as columns. Grade 11
## first, where the CT-PO deviance advantage over SBAC was largest.
X <- as.matrix(dt_g11[, .(ES, LS, PS)])

## 2. Pseudo-observations -----
## The rank transform. This is Sklar's theorem in one line: after it, the
## marginals are gone and what remains is purely the dependence structure.
## Ties broken at random per Genest et al. (2009) – with n in the hundreds of
## thousands and a discrete scale score, ties are not a rounding detail.
U <- pobs(X, ties.method = "random")
```

The fitting step is where the current assumption gets tested rather than inherited:

```
## 3. Fit the candidate families -----
## dispstr = "un" leaves all three pairwise correlations free. That is exactly
## the committee's April question – "would it make a difference if you didn't
## have the same correlation between the three?" – turned into a fitted
## parameter instead of an assumption.
fits <- list(
  gaussian = fitCopula(normalCopula(dim = 3, dispstr = "un"), U, method = "mpl"),
  t        = fitCopula(tCopula(dim = 3, dispstr = "un"),    U, method = "mpl"),
  frank    = fitCopula(frankCopula(dim = 3),                U, method = "mpl")
)

## Relative fit only. At Michigan's n every parametric family will be rejected
## by a formal goodness-of-fit test – that was true across all 966 conditions in
## the sensitivity study – so compare on information criteria and on fit in the
## tails, never on p-values.
data.table(
  family = names(fits),
  logLik = sapply(fits, logLik),
  AIC     = sapply(fits, AIC),
  BIC     = sapply(fits, BIC)
)[order(AIC)]
```

Then read off what the dependence actually looks like. The gaussian row above is the model the simulation currently assumes; the gap between it and the best row is the size of the assumption:

```
## 4. The structure, pair by pair -----
best <- fits$t
pairs <- apply(combn(colnames(X), 2), 2, paste, collapse = "-")

## Degrees of freedom: the t-copula's tail-dependence dial. Large df means it is
```

```

## behaving Gaussian and the family choice barely matters here. Small df means
## joint extremes are far more common than a normal implies – which is the case
## that would change the classification results.
coef(best)[["df"]]

## Everything else comes off the fitted trivariate object directly. getSigma()
## returns the correlation matrix; tau() and lambda() return one value per pair,
## ordered the same way as combn() above – (1,2), (1,3), (2,3).
##
## Read lambda_lower first. It is the probability of joint low performance, and
## the Below Standard cut sits exactly there. A Gaussian copula forces it to zero
## by construction, which is the assumption this whole exercise is testing.
P <- getSigma(best@copula)
lam <- lambda(best@copula)

data.table(
  pair      = pairs,
  rho       = P[lower.tri(P)],
  tau       = tau(best@copula),
  lambda_lower = lam[paste0("lower", seq_along(pairs))],
  lambda_upper = lam[paste0("upper", seq_along(pairs))]
)

```

And the substitution into the existing simulation, which is the part I want to stress is **small**:

```

## 5. Swap the dependence model, keep everything else -----
## The marginals do not change. They stay the fitted three-component mixtures
## from pages 5-7 of the deck. Only the copula is replaced. That is the whole
## practical value of the Sklar factorisation here: this is a substitution in
## the data-generating step, not a rewrite of the study.
##
## qgmm() is the inverse CDF of a fitted mixture – the marginal machinery
## already in the pipeline.
simulate_theta <- function(n, cop, gmm_by_subtest) {
  U <- rCopula(n, cop) # dependence
  vapply(seq_along(gmm_by_subtest), # marginals, untouched
    function(j) qgmm(U[, j], gmm_by_subtest[[j]]),
    numeric(n))
}

theta_t      <- simulate_theta(n_sim, best@copula, gmm_by_subtest)
theta_gaussian <- simulate_theta(n_sim, fits$gaussian@copula, gmm_by_subtest)

## Then run the existing CT-P0 and SBAC classification on each and compare QWK.
## Two informative outcomes, no uninformative one:
##

```

```
## numbers agree -> the dependence model is not what limits classification.  
##           The ceiling is unidimensionality, and the reporting  
##           question is the whole game.  
## numbers differ -> Batch 1 was run at a dependence structure the data does  
##           not exhibit, and Batch 3 should be specified from the fit.
```

Everything above runs on data already collected, and the only new dependency is the copula package. If it would help, I am happy to run the first pass myself against whatever extract is straightforward to share.

A second place the same machinery applies

Worth noting because it connects two sessions that did not talk to each other in April.

Michigan has **Smarter Balanced interim and M-STEP scores on many of the same students** — that is what makes [M-STEP / Smarter Balanced Concordance](#) possible at all. The concordance table maps a performance-level band on one assessment to a band on the other. That is marginal-to-marginal alignment, and it discards the joint distribution entirely.

The committee pressed on the word *could* in the table's worked example — “a 3rd grade student scoring 2112 on the interim **could** score 1203 on the M-STEP” — and got “would most likely score, or something along those lines.” A copula fitted to the (interim, M-STEP) pair gives the actual conditional distribution of the M-STEP score given an interim score. That turns “could” into a statement with a quantified spread, which is what [Consider a prediction study if participation supports it](#) is asking for and what the table's own disclaimer — “not a perfect prediction” — currently gestures at without measuring.

Same machinery, different pair, data already collected.

What I am asking

Not that anyone adopt a method because I suggested it. Three things:

1. **State the research question on the first slide.** My honest experience in April was of following each analysis and losing the thread of what they were collectively for. If Q3 is the question, the Batch 2 gradient is the finding and everything else is support.
2. **Name the dependence assumption.** A correlation matrix on latent normals is a Gaussian copula, and it is the family that fits educational data least often. Naming it makes it checkable.
3. **Run the bounded check above,** or tell me why it is the wrong check. Either answer is progress.

I offered in April to work on this with Ji Zeng ([Support Ji Zeng's work through direct collaboration](#)). That offer stands, and this is the concrete form of it.

Related

M-STEP Subscore Reporting · CT-PO vs SBAC Simulation Results Deck · Model the joint density of the three dimensions · Establish scale, then cuts, then categorise from theta · Review item parameters and total-score correlations before Batch 3 · Would varying the inter-subtest correlations change the result? · M-STEP / Smarter Balanced Concordance · Ji Zeng · Damian Betebenner