

The maths behind the projection

Which statistical model we fit, why we chose it over the obvious alternatives, and what it is still wrong about.

Written for a Head of Mathematics who intends to argue with it.

Version 1.0 (public release) · 17 September 2026 · Review: March 2027

A projection engine is a sequence of choices, and most of them can be got wrong in ways that are invisible from the output. This paper sets out each choice, the alternative we rejected, and the reason. Where a method is assumed rather than measured, it says so. Where we found our own work to be wrong, it says that too.

01

The problem, stated precisely

A school holds assessment marks for students part-way through a course. It wants, for each student, a projected final HSC mark in each subject and a projected ATAR, with an honest statement of how uncertain both are.

Three features of the HSC make this harder than a standard regression problem.

The quantity that matters is a rank, not a mark. An HSC mark is moderated against a state-wide cohort, and the ATAR is a rank derived from scaled marks. A student's raw school mark is not on the same scale as the thing being predicted, and the relationship between them is set by the strength of the cohort they sit in.

Cohorts are small. A single school's Extension 2 Mathematics class may have eight students. Any method that needs a large sample per subject per year will not survive contact with real school data.

The target moves. NESA rewrites syllabuses on a rolling schedule. Results under a superseded syllabus describe a different examination from the one this year's students will sit.

WHAT THIS RULES OUT IMMEDIATELY

Any model that predicts a raw mark directly from prior raw marks, without passing through rank, is fitting the wrong quantity. It will look accurate in a stable cohort and fail in the year a strong group moves through.

02

Why rank, and why isotonic regression

The engine's core object is a curve, fitted per subject, mapping a student's **rank-fraction** within their school cohort (0 at the top, 1 at the bottom) to a predicted final HSC mark. It is fitted on that school's own historical results: for each past student we know where they ranked internally and what they finished with.

Rank is the right input because it is the quantity a school's internal assessment actually measures well. Teachers rank a cohort reliably; they calibrate absolute marks against a state standard far less reliably, because they cannot see the state.

The curve must be monotone, and it must not be a straight line

Two properties are non-negotiable. It must be **non-increasing**: ranking higher cannot predict a lower mark. And it must be free to **bend**, because the relationship genuinely is not linear. Marks compress near the top of the range and spread out in the middle.

We fit it with **isotonic regression** using the pool-adjacent-violators algorithm. It finds the monotone curve minimising weighted squared error, with no functional form assumed.

ALTERNATIVE	WHY WE DID NOT USE IT
Ordinary least squares	Assumes a straight line the data does not follow, and can produce a non-monotone fit on small samples, predicting a lower mark for a higher rank.
Polynomial or spline fit	Bends, but is not constrained to be monotone. On a cohort of twelve, a cubic will happily wiggle and invent structure that is noise.
Quantile regression	A reasonable choice, and it would give band estimates directly. It needs more data per subject than a school HSC cohort provides, and its bands are themselves noisy at these sample sizes.
Hierarchical / mixed-effects model	The statistically ideal answer if we had the whole state. We do not: each school's data is its own, and a model that pools across schools is one we cannot fit and a school cannot audit.

THE HONEST LIMITATION OF ISOTONIC REGRESSION

The fit is a step function, which we interpolate linearly between control points. Its known weakness is at the boundary: the monotonicity constraint is one-sided there, the estimator is not consistent at the endpoints, and the pool-adjacent-violators step can absorb a wide range of ranks into a single flat block. This is why the backtest reports error by rank decile rather than as a single average.

03

Old years still count, but less

A school's history is not uniformly relevant. A result from six years ago describes teaching, students and often a syllabus that have all moved on.

The alternative approaches are a hard cut-off, which throws away sample a small cohort cannot spare, and using everything equally, which lets stale data outvote current data. Both are worse than a weight that declines smoothly.

We weight each historical result by **exponential decay on its age**, on a half-life chosen to reflect how quickly a school's senior teaching turns over. A result from several years ago still informs the curve; it simply counts for less than this year's.

The half-life is **pre-registered, not fitted**: chosen in advance on that reasoning, and not tuned to improve backtest numbers. Tuning a parameter on the same data you evaluate on is how a model comes to look better than it is, and it is the single easiest way for a vendor to produce an impressive accuracy figure that means nothing.

04

Syllabus changes are a hard boundary, not a weight

Recency weighting handles gradual drift. A syllabus rewrite is not drift: it is a different examination, and results from before it are not weaker evidence but evidence about something else.

The engine buckets calibration by **subject and syllabus version**. Old-syllabus and new-syllabus results are never pooled. A subject whose syllabus has just reset therefore has little or no usable history, however long the school has been teaching it, and the interface says so rather than quietly borrowing from the old course.

The data tier shown against every projection takes the *worse* of two constraints: what the syllabus permits, and what this school has actually connected. A school with seven years of its own results in a subject NESA reset last year is still, correctly, told the projection is not tuned to them yet.

05

From marks to an ATAR

Scaling is not modelled, and it is not guessed. UAC computes the scaling relationship from the cohort that has just sat the examination and publishes it afterwards, so it is necessarily retrospective: a projection made today applies the most recently published tables, which describe last year's candidature. A subject whose candidature shifts will scale differently, and that difference sits **outside** the confidence band described in Section 06, which propagates mark uncertainty only. We would rather name that than let a reader assume the band covers it. The engine applies the published tables: the engine converts a projected HSC mark into scaled marks per unit for that subject.

The aggregate is the sum of the student's best **10 units** by scaled mark per unit — but the selection is constrained, and the constraints are the part that is easy to get wrong. The UAC pattern requires **2 units of English** in the aggregate, permits at most **2 units of Category B**, and counts at most **7 units of Science**.

The English requirement is the one that changes numbers. A strong mathematics and science candidate with a weak English mark has English ranked last, so an unconstrained "best ten" drops it — and returns a figure that is not an ATAR at all. On a realistic profile of that shape the unconstrained sum overstated the result by **3.5 ATAR points**. It is also the first case a Head of Mathematics is likely to test, because every school has that student.

Two smaller details in the same step. The comparison is on **per-unit** value, not the product of value and units, so a one-unit Extension course at a high per-unit value can outrank a two-unit course at a lower one. And the ordering is given an explicit tie-break, because two students with identical marks must receive identical counting sets.

The aggregate then converts to an ATAR through UAC's published aggregate-to-ATAR mapping. The relationship is steep in the middle and very flat near 99.95, which is why the band around a projected ATAR is **asymmetric**: we push the aggregate's lower and upper bounds through the mapping separately rather than reporting a symmetric plus-or-minus.

06

The confidence band

A projection without a band is a claim of certainty nobody has. Ours is built by propagating per-unit uncertainty into the aggregate, then through the ATAR mapping.

Each counted unit carries a standard deviation set by its **data tier**: tightest where the school's own same-syllabus curve backs the projection, wider where history is thin, widest where the projection leans on public scaling and live rank alone. Each is then scaled by the local slope of that subject's scaling curve, because one mark buys different amounts of scaled aggregate in a high-scaling subject than a low-scaling one.

Errors are correlated, and pretending otherwise would understate the band

Adding variances independently would be the convenient assumption and the wrong one. A student having a poor year tends to have a poor year across subjects, so the errors do not cancel the way independence assumes. On a typical counted load, treating the different courses as independent while still counting the two units of one course as a single measurement produces a band about a third narrower.

Treating every unit as independent — the naive sum of variances — makes it roughly half. Either way the error runs in the direction that flatters the vendor. The engine uses a **block correlation model**: the two units of one course are effectively one measurement, so their correlation is 1; different courses share a student-and-cohort factor at a moderate positive correlation.

THIS IS AN ASSUMPTION, AND IT IS STATED AS ONE

That correlation is pre-registered, not measured, and we report its sensitivity across a range either side of the chosen value so a reader can see how much the band depends on it. The tier standard deviations are likewise assumed until a school's own backtest replaces them with measured residual spread. Until then the band is **indicative and labelled as such**, not sold as a measured prediction interval. Every band the interface draws carries its provenance for this reason, and we will show a school the sensitivity figures on request.

07

A bias we found in our own engine

This section describes a defect in our own work, because how a vendor handles finding one is more informative than any accuracy claim.

The best-10 rule **selects on the same projections it then sums**. A subject projected above its true value is both more likely to be selected and contributes that upward error to the total; one projected below is more likely to be dropped before it can contribute the offsetting downward error. The errors do not cancel, so every projected ATAR was optimistic. Giving the reporting step an independent second noise draw in simulation collapses the bias from **+0.91 ATAR to -0.03**, which identifies the shared draw as the whole of the cause.

The obvious fix does not work, and we can show why

The textbook response is shrinkage: pull the values the optimiser ranks on toward a central value, so a subject that only looks best because of noise stops winning. We had implemented exactly that. It removes **none** of the bias, because pulling every value toward a common centre is a monotone transformation and preserves the ordering. Measured across four student profiles, three history depths and two noise levels, it removed 0% in every cell.

Genuine empirical-Bayes shrinkage uses a per-subject reliability and centre, which is not order-preserving and would change the selection. Estimating those needs a pool of schools we do not have and would not want to require.

What does work

The correction has to act on the reported total. The engine estimates the optimism by a **plug-in Monte Carlo procedure**: treat the observed per-unit values as a stand-in for the truth, re-run the exact select-then-sum pipeline many times over with fresh noise at each unit's own standard deviation, and average the inflation that produces. That average is subtracted from the aggregate before conversion to an ATAR.

STUDENT PROFILE	BIAS BEFORE	BIAS AFTER	MAE BEFORE	MAE AFTER
Mid, even spread	+0.875	+0.152	1.700	1.612
High, even spread	+0.517	+0.057	1.084	1.066
Narrow spread	+1.065	+0.296	1.729	1.560
Wide spread	+0.277	-0.056	1.575	1.608

Between 72% and 90% of the bias removed, with mean absolute error improving in three profiles of four. No displayed subject mark moves and the counting selection is unchanged; only the aggregate is corrected, and only downward.

READ THESE FIGURES FOR WHAT THEY ARE

They come from simulation, and the correction assumes the same noise model the simulation generates from. They establish that the method is **internally consistent under its own assumptions**, not that it performs this well on real results. The residual is roughly +0.15 in the middle case, and we report it rather than round it away.

08

How we know any of this works

On a school's connected history the engine runs **leave-one-year-out** backtesting. For each historical year it refits calibration on the remaining years, predicts every student in the held-out year, and compares against what actually happened.

Folds are by **year**, not by student. A cohort is the correlated block: it shares teachers, cohort strength and exam conditions. Holding out random students would leak those shared conditions into the training set and report an error far lower than the engine will achieve on a year it has never seen.

It reports mean absolute error, root mean squared error, signed bias, the empirical error standard deviation, and **coverage**: the share of results falling inside the stated band. The interval around that coverage figure is a **Wilson score interval**, not the more familiar Wald one, because at the three to six years a school typically connects the Wald interval returns upper bounds above 1.0 and collapses to zero width when every held-out year happens to fall inside the band. It is computed on the number of independent year-blocks rather than students: five years of data is five independent observations of the engine, not six hundred.

The confound we have to design around

An HSC mark is a 50/50 blend of the external examination and the school's own moderated internal assessment, and our predictor is the student's rank, derived from that same internal assessment. So part of the relationship is **arithmetic rather than prediction**. Any engine predicting HSC marks from school assessment has this problem; most will report the flattering figure without mentioning it.

We validate against **both** targets. The external examination is marked by NESA blind to the school, so a rank-to-exam relationship carries no such leak and is the honest test. The backtest reports the ratio between the two residual spreads. That ratio is always above 1, because an examination carries noise no internal assessment can see; what matters is how far above. We continue to present the HSC mark, because that is what appears on the certificate, but we calibrate knowing the difference and a school can see both figures.

09

Not every pattern is a finding

Surfacing patterns from small cohorts invites false positives. A class of eight will produce apparent effects by chance alone, and a platform that flags them all trains its users to ignore it.

Before any pattern is raised, the engine runs an **exact two-sided binomial test** against the relevant null — exact rather than the normal approximation, which over-flags at precisely the small sample sizes the guardrail exists to protect. Across the many patterns evaluated for one school, the **Benjamini-Hochberg** procedure controls the false-discovery rate, so the volume of tests does not manufacture spurious flags.

Benjamini–Hochberg controls that rate under independence or positive dependence. The tests here are computed on the same students across different topics, so a weak cohort depresses several at once — which is the positive-dependence case the procedure is valid for. We state the condition rather than leave it implicit, because it is the first thing a statistician checks.

10

What we do not claim

An engine's limitations belong in the same document as its method.

- **We publish no accuracy percentage.** Any figure quoted before a school's own history is connected would describe our test data, not their students. The backtest produces a school's own numbers, and those are the ones that matter.
- **The band is indicative until calibrated.** The tier standard deviations and the cross-subject correlation are pre-registered assumptions, not measurements, until a school's backtest replaces them.
- **The ends of the curve are the weakest part.** Fewest students, most extrapolation. The per-decile error breakdown exists to keep that visible.
- **Selection bias is reduced, not eliminated.** Roughly +0.15 ATAR remains in the middle case.
- **Scaling is retrospective.** A projection made today applies the most recently published tables, which describe last year's candidature. That error sits outside the confidence band.
- **We model students on an ATAR pathway.** A projection says nothing about a student for whom that is not the relevant question, and a substantial minority of each cohort does not receive a tertiary entrance rank at all.
- **A projection is a signal, not a verdict.** It says where attention is worth spending now. It is not a ceiling, not a target, and not a statement about what a student is capable of.

ON WHAT THIS PAPER DOES AND DOES NOT SPECIFY

This document describes the method in full: which model is fitted, which alternatives were rejected and why, how uncertainty is propagated, and how the engine is tested. It does not print the specific tuning constants — the decay half-life, the assumed correlation, the per-tier standard deviations. Those are implementation parameters, and the techniques they sit inside are standard published statistics rather than anything we invented.

They are not withheld from customers. A school evaluating Avantus, or an academic reviewing the method, can have the parameters and the sensitivity analysis on request. We would rather have that conversation with someone accountable for the decision than publish a configuration file to the open web.