

WHITE PAPER

How the engine is tested

A projection is only worth what the testing behind it is worth. This is what we test, what the testing has found, and what we changed as a result.

Version 1.0 · 17 September 2026 · Review: March 2027

CONTENTS

- | | |
|---|--|
| 01 Why publish this at all | 04 The fix that did not work |
| 02 Correlated errors, and a band that was too narrow | 05 An interval that produced impossible numbers |
| 03 A selection effect nobody would notice | 06 A review briefed to find errors |
| | 07 What we have not solved |

Any vendor can tell you their model is accurate. Far fewer will tell you what they went looking for, what they found, and what it cost them to fix. This paper is the second kind of document. It is a record of the questions we put to our own engine and the answers we got back, including the ones we would rather have not.

01

Why publish this at all

A school buying a projection engine is being asked to trust a number it cannot check. The student has not sat the examination yet, so there is no result to compare against, and by the time there is, the decisions that mattered have already been made.

That puts the burden on the vendor to show its work. Not a claimed accuracy figure, which is unfalsifiable before the fact and easy to produce from favourable test conditions, but the testing itself: what we suspected might be wrong, how we measured it, and what the measurement said.

We would rather be the company that publishes what its own testing found than the one that publishes a number.

02

Correlated errors, and a band that was too narrow

What we suspected. The confidence range around a projected result is built by combining the uncertainty in each counted course. The convenient way to combine them is to treat each course as an independent measurement, and that is what a first implementation naturally does.

Why it is wrong. Students do not work that way. A student having a strong run tends to have it across several courses at once, and a student having a difficult year tends to have it broadly. When errors move together they do not cancel the way independence assumes, and the resulting range is narrower than the evidence supports.

What we changed. The engine now models the correlation explicitly. Two units of the same course are treated as a single measurement, because that is effectively what they are, and separate courses are given a moderate positive correlation. The ranges this produces are visibly wider than the independent calculation gives.

WHAT THIS MEANS FOR A SCHOOL

A wider range is the less flattering result, and it is the honest one. A range that is too narrow invites a school to act on a distinction the data cannot actually support.

03

A selection effect nobody would notice

What we suspected. An ATAR aggregate counts a student's best ten units. When those units are chosen using projections that each carry uncertainty, the ones that look strongest are slightly more likely to be the ones sitting high by chance rather than by ability.

Why it matters. Selecting on that basis and then reporting the same values nudges the total upward. The errors do not cancel, because the selection step and the reporting step share the same uncertainty. It is a small effect, and it runs in the direction that flatters the vendor, which is precisely why we went looking for it.

What we changed. The engine estimates the size of that optimism by re-running the whole selection process many times under the uncertainty each course carries, then subtracts it. No result shown to a student moves, and which courses count is unchanged. Only the total is corrected, and only downward.

04

The fix that did not work

This is the finding we are most glad to have made.

There is a standard response to the problem in Section 03: adjust the values used to make the selection, so that a course which only looks strongest because of chance no longer wins on chance. We built it. It was well reasoned, carefully commented, and covered by tests.

It removed none of the effect.

The reason is simple once seen. The adjustment moved every value by a proportional amount toward a common centre, which leaves their order unchanged. The selection is a ranking, so a change that cannot alter the ranking cannot alter what gets selected. We measured it across a range of student profiles and levels of uncertainty, and the result was the same in every case.

Knowing why an obvious answer fails is worth as much as the answer. It sent us to a different approach, acting on the total rather than the ranking, which does work.

We deleted the original code rather than leave it in place. Something that looks like a correction but provably is not is worse than nothing, because the next person to read it assumes the problem is handled.

05

An interval that produced impossible numbers

What we suspected. The engine reports how often past results actually fell inside the range it gave, and puts a confidence interval around that figure. Schools connect a handful of years of history, so that interval is computed on very few independent observations.

What we found. The standard textbook interval misbehaves badly at those sample sizes. It returned upper bounds above one hundred per cent, which is not a possible value, and where every year happened to fall inside the range it collapsed to zero width, claiming certainty from four observations.

What we changed. A different interval, chosen specifically because it stays valid at small sample sizes. Where the old calculation claimed certainty, the new one reports a range beginning near fifty per cent, which is the honest statement that four clean years do not establish that a range is right.

06

A review briefed to find errors

Confirming your own work is the easy test to pass, so before publishing our methodology we ran a separate adversarial review across the whole engine. The brief was to find errors rather than to confirm the work: to check every claim in the draft against the actual source code, and to report what a hostile expert reader would attack first.

It found more. Several were places where the statistics were described slightly wrong rather than done wrong, which matters in a document whose purpose is to be checkable. Two were real defects in the engine, fixed and tested the same day. All of it was corrected before anything was published, and the review is the reason this paper exists in its current form.

We mention the process rather than the finding count because the process is the part a school can hold us to. Any vendor can say their engine was reviewed. The question worth asking is whether the review was briefed to agree.

THE LIMITATION OF THIS DOCUMENT

Everything above is testing against simulated students and our own historical data. The figure that will matter to your school is the one produced by running this process against **your** results, which is what the engine does once a school connects its history. Until then, treat our testing as evidence of method rather than evidence of accuracy.

07

What we have not solved

- **The correction in Section 03 is substantial, not complete.** A small amount of optimism remains, and we report it rather than round it away.
- **Scaling is retrospective.** The published scaling relationship describes the cohort that has already sat the examination, so a projection made today uses last year's. A course whose candidature shifts will behave differently, and our range does not currently model that.
- **The ends of the curve are the weakest part.** Projections for students at the very top and very bottom of a cohort rest on the least evidence, which is why our testing reports error separately for each tenth of a cohort rather than as a single average.
- **We publish no accuracy percentage.** Any figure quoted before a school connects its own results would describe our test conditions rather than their students.

This list will change. It will not disappear, because a projection engine that has stopped finding things wrong with itself has stopped being tested.