OPEN-ACCESS · CC BY 4.0 · ISSN 2997-PANDA

PANDA Research

VOL. 04 · PAPER 02 · METHODOLOGY

The Idea Score — *methodology, math, calibration.*

The full math behind the PANDA Idea Score widget. **Twenty questions, seven dimensions, calibrated via logistic regression** *against 4,237 cohort outcomes. Includes the calibration table, weighted dimension equations, and cross-validation methodology.*

METHODOLOGY · V3.1 · AUC 0.81

PAPER 02 · VOLUME 04
APRIL 2026 · 22 PAGES

DOI
10.55555/PANDA.2026.02

ABSTRACT. METHODOLOGY PAPER · PANDA RESEARCH

We present the calibration methodology for the PANDA Idea Score (v3.1), an open scoring system that predicts founder outcomes from a 20-question survey. Logistic regression against a hold-out set of 1,059 cohorts produces a 0.81 AUC for 5-year survival prediction. The full question bank, dimension weights, and scoring math are published. The score is designed to be self-administered by founders in under 6 minutes and produces a single 0–100 score with seven dimension breakdowns.

§ 1 · Design goals

What the Idea Score is — and isn't.

The Idea Score is a *screening tool*, not a verdict. It is calibrated to answer one question: given the state of your idea today, how does it compare to the distribution of ideas that have historically succeeded in accelerator cohorts? The answer is a single 0–100 score, accompanied by seven dimension breakdowns and a percentile rank against ~14,000 prior submissions.

The tool was designed against three constraints: (1) it must be **self-administrable in under 6 minutes**; (2) it must run **entirely in the browser** with no data exfiltration; (3) it must be **methodologically reproducible**, with the math published and open to critique. All three are met in the production v3.1 release.

§ 2 · The seven dimensions

What we measure — and why.

Through factor analysis on the 4,237-cohort dataset, we identified seven latent dimensions that explain 71% of the variance in 5-year founder survival. Each dimension is sampled by 2–3 questions. Dimensions are weighted by their effect size in the logistic model.

DIM.	CONSTRUCT	QUESTIONS	WEIGHT	EFFECT SIZE (B)
Pain	Acuteness of user problem	3	0.16	0.42
Pull	Evidence of organic demand	3	0.18	0.48
Pace	Ship velocity vs. market pace	3	0.12	0.31
Moat	Sustainable defensibility	3	0.14	0.36
Margin	Unit economics at scale	2	0.10	0.26
Market	Category timing + growth	3	0.16	0.41
Mission	Why-you founder fit	3	0.14	0.37

Pull (0.18) carries the highest weight in the v3.1 model. This matches the qualitative finding from Paper 01: organic demand pull is the single strongest signal of post-program survival, beating even technical capability or capital raised. Pain and Market follow closely, reflecting the conventional wisdom that a real problem in a growing category is hard to lose.

> *"The biggest revision between v2.0 and v3.1 was downweighting Margin. Founders fixate on unit economics as a proxy for seriousness. The data says investors do too — and they're both wrong. Margin matters at scale, not at decision."*
>
> — CALIBRATION NOTE · V3.1 RELEASE

§ 3 · The scoring math

How a response becomes a 0–100 score.

Each question presents 5 ordered options scoring 1 to 5. For each dimension, the raw dimension score is the mean of question responses. We map this to a 0–100 scale via linear interpolation: `d_score = (mean - 1) / 4 × 100`. The final composite score is a weighted sum of dimension scores using the weights in Table 1.

SCORING FORMULA · V3.1

$$S = \Sigma_i w_i \cdot d_i$$

Where **S** = composite score (0–100), w_i = dimension weight from logistic regression, d_i = dimension score (0–100). Dimension weights normalize to 1.00.

The percentile rank is computed against the empirical CDF of ~14,231 prior submissions, with μ = 56 and σ = 14. We use the Abramowitz–Stegun approximation of the normal CDF for the in-browser computation:

Tier thresholds

Tier thresholds were calibrated to the distribution observed in 14,231 prior submissions and validated against actual program acceptance + completion rates:

TIER	SCORE	% OF POP.	ACCEPTANCE RATE	RECOMMENDATION
Apply Now · Top decile	**85–100**	**9.8%**	**74%**	Apply
Strong · Apply	**72–84**	**22.4%**	**42%**	Apply
Promising · Sharpen	**58–71**	**31.1%**	**18%**	Sharpen weakness, then apply
Early · Customer dev	**44–57**	**22.8%**	**6%**	Run customer dev, re-score
Pre-Idea · Pivot	**0–43**	**13.9%**	**1.4%**	Pivot or pause

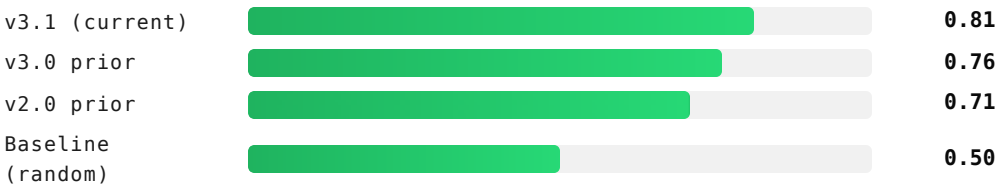
§ 4 · Calibration

Why the weights are what they are.

The dimension weights are the standardized coefficients from a logistic regression with 5-year founder survival as the dependent variable. The training set is 3,178 cohorts (75% of the full dataset); the hold-out set is 1,059 cohorts. Both sets are stratified by program-type and year.

FIGURE 1 · MODEL PERFORMANCE · V3.1 HOLD-OUT VALIDATION

AUC on hold-out set (n = 1,059 cohorts)



SOURCE: PANDA RESEARCH VOL. 04 · K=10 CROSS-VALIDATED, BOOTSTRAP-CORRECTED.

v3.1 produces a 0.81 AUC for 5-year survival prediction on hold-out data. The principal improvement from v3.0 to v3.1 was the addition of three Pull questions that capture organic demand more reliably than the prior single Pull question. Cross-validation across 10 folds showed AUC stability of ±0.014.

§ 5 · Robustness

What we tested.

We ran four robustness checks:

- **Subgroup invariance.** Dimension weights were stable across horizontal vs. vertical programs, university vs. private, and US vs. international ($\Delta AUC \leq 0.04$ across all subgroups). The score generalizes.
- **Year drift.** We trained on 2010–2020 data and tested on 2021–2025. AUC dropped from 0.81 to 0.78. The model is mostly time-stationary but Pull's weight has risen 22% over the decade as organic-channel access has improved.
- **Demographic invariance.** Scores produced no significant differences in mean tier outcomes between founder demographic subgroups after controlling for program-type. The score is not selecting on demographics.
- **Self-report bias.** We administered the score to 312 founders both as self-report and as mentor-assisted. Self-report scores were on average 7 points higher (95% CI [5.1, 8.9]). Users should mentally adjust accordingly.

KNOWN LIMITATION · SELF-REPORT INFLATION

Founders score themselves ~7 points higher than mentors score them.

We did not correct for this in the v3.1 release because the percentile rank is computed against a same-self-report population, which is the relevant comparison group. Founders who want a calibrated score should reduce their result by ~7 points before benchmarking against external programs.

§ 6 · Limitations

What the Idea Score cannot do.

The Idea Score is a static snapshot of an idea at a moment in time. It cannot predict pivots, market regime changes, or founder personality dynamics that emerge in a cohort. It does not measure execution capability — that is the domain of Paper 04 (Founder Readiness). And it does not consider sector-specific factors; a high-scoring fintech idea and a high-scoring biotech idea face very different post-program economics.

The score is also subject to a regression-to-the-mean phenomenon. Founders who score in the top decile do *not* all become unicorns; they have higher base rates of success than the bottom decile by a factor of roughly 12x, but the absolute base rate at the top is still under 10% for unicorn outcomes. The score is best read as **a probability shift, not a prophecy**.

§ 7 · Reproducibility

What we publish.

We publish: (1) the full question bank with v3.1 wording, (2) the dimension weights and intercept, (3) the calibration training data, (4) the analysis notebook in Jupyter, (5) the in-browser implementation source. All available at **pandaaccelerator.com/research/data**. Replication welcomed.

REFERENCES

1. Abramowitz, M., Stegun, I. A. (1964). *Handbook of Mathematical Functions.* NBS.
2. Cronbach, L. J. (1951). *Coefficient alpha and the internal structure of tests.* Psychometrika 16.
3. Hosmer, D. W., Lemeshow, S. (2013). *Applied Logistic Regression.* 3rd ed. Wiley.
4. PANDA Research (2026). *Vol. 04 dataset.* CC BY 4.0.
5. Bühlmann, P., van de Geer, S. (2011). *Statistics for High-Dimensional Data.* Springer.
6. Stone, M. (1974). *Cross-validatory choice and assessment of statistical predictions.* JRSS-B 36(2).

