

Tracking Performance Gains in AI-Supported Learning:

A Longitudinal Study of Criterion-Referenced AI Feedback on StudyPulse

StudyPulse Research Team

November 2025

Abstract

We present a longitudinal analysis of 28,279 graded submissions from StudyPulse users (February–November 2025), all on authentic Victorian Curriculum and Assessment Authority (VCAA) past examination questions spanning 7 years of real VCE exams (2018–2024). Performance is difficulty-adjusted using z -score normalisation against official VCAA score distributions, which represent cohorts of 50,000+ students per year. By the end of the study window, StudyPulse users achieve a mean performance level of approximately 0.7–0.8 standard deviations above the exam population average, corresponding to roughly +5 VCE study score points (from 30 to ≈ 35) and +10 ATAR points (from 70 to ≈ 80). Within the observation period, we observe a mean improvement of about +0.55 standard deviations from early- to late-year performance. Under a normal-distribution model, we further estimate that StudyPulse users are approximately 3 \times more likely to achieve a 40+ study score compared to the general student population ($\approx 23\%$ vs. $\approx 8\%$).

Keywords: Educational Technology, AI Tutoring, Z-Score Analysis, VCE, Learning Analytics

1 Introduction

1.1 Background

The Victorian Certificate of Education (VCE) uses a standardised scoring system in which study scores are scaled to a population mean of 30 and standard deviation of 7. This normalisation allows for meaningful comparison across subjects and years. Similarly, the Australian Tertiary Admission Rank (ATAR) is typically modelled as an approximately normal distribution for the ATAR-eligible cohort, with mean around 70 and standard deviation around 14.

In this work we exploit these scaling conventions to place StudyPulse users' performance on a difficulty-adjusted scale that is directly comparable to population-level exam performance.

1.2 Why Authentic VCAA Questions Matter

This study is based exclusively on *real* VCAA past examination questions from the last 7 years of VCE exams (2018–2024). This choice is critical for three reasons:

1. **Validity.** Students are practicing on the exact question formats, rubrics, and difficulty levels they will encounter in actual exams.

2. **Reliable benchmarks.** VCAA publishes official score distributions for each question, representing 50,000+ real student responses per year, which provide a robust population benchmark.
3. **No synthetic inflation.** Unlike practice questions or third-party materials, VCAA questions have known and fixed difficulty profiles that cannot be artificially adjusted to make gains appear larger than they are.

1.3 Why Immediate Feedback Matters

The value of instant, criterion-referenced feedback is well-established in learning science. Hattie and Timperley’s synthesis of over 800 meta-analyses identifies feedback as one of the highest-impact influences on achievement, with an effect size of $d \approx 0.7$, nearly double the average effect of a year of schooling [4]. Crucially, feedback must be *timely*: when corrections arrive days or weeks after the student has forgotten their reasoning, the opportunity to reshape underlying skills has largely passed.

Neuroimaging research supports this: immediate feedback engages striatal systems associated with procedural learning and skill automaticity, whereas delayed feedback shifts learning toward declarative memory systems better suited to memorising facts than tuning exam technique [5].

Recent data quantifies the behavioural difference. In a study of approximately 19,500 Australian and New Zealand secondary students, Education Perfect found that when AI feedback was immediate, **83%** of students with a low-quality response attempted a revision; without immediate feedback, only **7%** did [6]. Instant feedback transforms assessment from a one-shot judgment into an iterative learning loop.

StudyPulse is designed to deliver this high-frequency, criterion-referenced feedback at scale, not as a generic chatbot, but as a strict, curriculum-aligned examiner that references official VCAA rubrics and marking schemes.

1.4 Research Questions

We investigate the following research questions:

1. How do StudyPulse users perform relative to exam population norms when controlling for question difficulty?
2. What is the magnitude of improvement in terms of VCE study scores and ATAR equivalents?
3. What is the probability that a StudyPulse user attains a high study score (40+) under reasonable distributional assumptions?

2 Data and Methodology

2.1 Data Collection

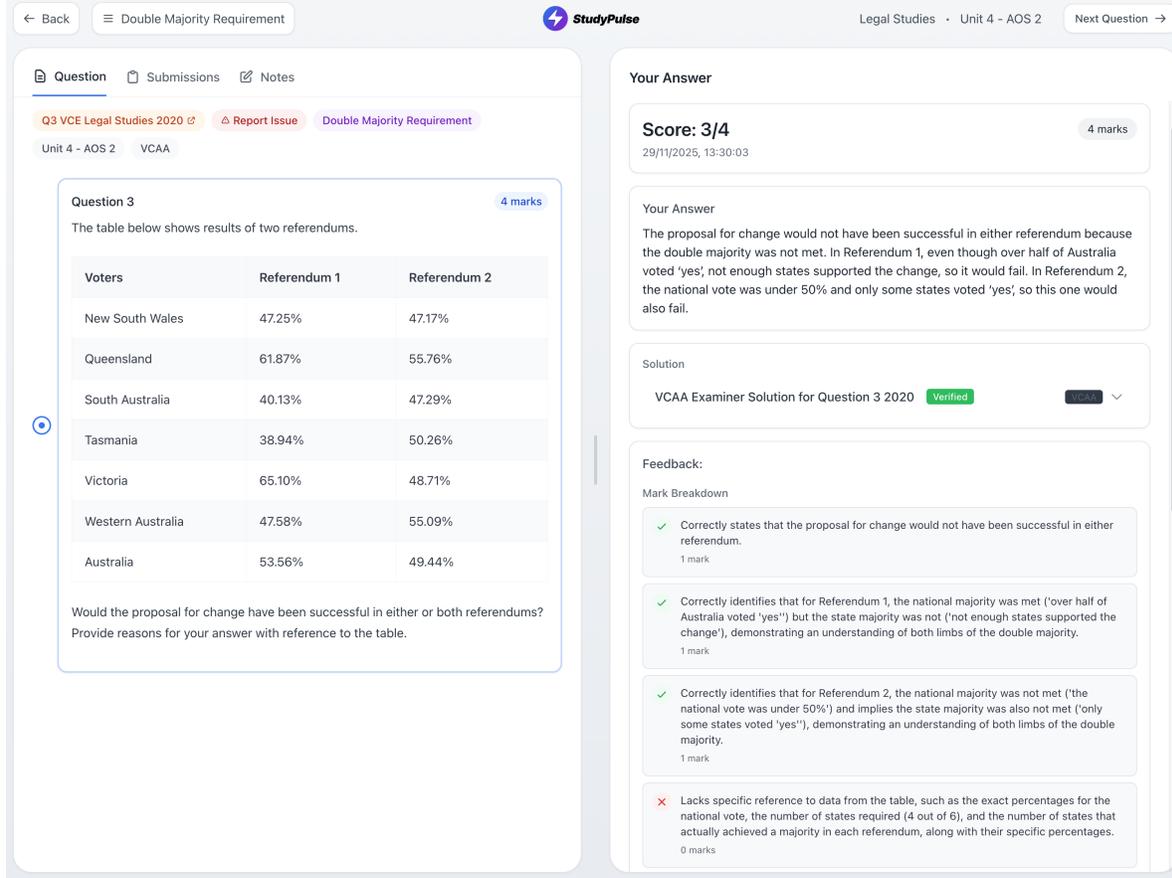
We analyse graded submissions generated by students using the StudyPulse platform.

- **Sample size:** 28,279 graded submissions.
- **Question source:** Authentic VCAA past examination questions only (no third-party or synthetic questions).

- **Exam years covered:** 5 years of VCE examinations (2018–2023).
- **Time period:** 10 February 2025 – 17 November 2025 (student submission window).
- **Subjects (11 VCE subjects):**
 - Physical Education (PE)
 - Legal Studies
 - Business Management
 - Health and Human Development (HHD)
 - Psychology
 - Geography
 - Economics
 - Religion & Society
 - Accounting
 - Biology
 - Physics

Each submission corresponds to a student’s attempt on a single exam question for which official VCAA marking schemes and score distributions are available.

Figure 1 illustrates the StudyPulse feedback interface. A student has answered Question 3 from the 2020 VCE Legal Studies exam, a 4-mark question requiring analysis of referendum voting data using the double majority rule. Within seconds of submission, the platform delivers criterion-referenced feedback: 3 marks are awarded with specific justifications tied to each rubric criterion, while 1 mark is withheld with an explanation that the student’s response lacked direct reference to the table data (exact percentages, number of states required, etc.). This immediate, granular feedback exemplifies the “teachable moment” that traditional delayed marking cannot provide.



← Back Double Majority Requirement StudyPulse Legal Studies · Unit 4 - AOS 2 Next Question →

Question Submissions Notes

Q3 VCE Legal Studies 2020 Report Issue Double Majority Requirement

Unit 4 - AOS 2 VCAA

Question 3 4 marks

The table below shows results of two referendums.

Voters	Referendum 1	Referendum 2
New South Wales	47.25%	47.17%
Queensland	61.87%	55.76%
South Australia	40.13%	47.29%
Tasmania	38.94%	50.26%
Victoria	65.10%	48.71%
Western Australia	47.58%	55.09%
Australia	53.56%	49.44%

Would the proposal for change have been successful in either or both referendums? Provide reasons for your answer with reference to the table.

Your Answer

Score: 3/4 4 marks
29/11/2025, 13:30:03

Your Answer

The proposal for change would not have been successful in either referendum because the double majority was not met. In Referendum 1, even though over half of Australia voted 'yes', not enough states supported the change, so it would fail. In Referendum 2, the national vote was under 50% and only some states voted 'yes', so this one would also fail.

Solution

VCAA Examiner Solution for Question 3 2020 Verified ⌵

Feedback:

Mark Breakdown

- ✓ Correctly states that the proposal for change would not have been successful in either referendum. 1 mark
- ✓ Correctly identifies that for Referendum 1, the national majority was met ('over half of Australia voted 'yes') but the state majority was not ('not enough states supported the change'), demonstrating an understanding of both limbs of the double majority. 1 mark
- ✓ Correctly identifies that for Referendum 2, the national majority was not met ('the national vote was under 50%') and implies the state majority was also not met ('only some states voted 'yes)'), demonstrating an understanding of both limbs of the double majority. 1 mark
- ✗ Lacks specific reference to data from the table, such as the exact percentages for the national vote, the number of states required (4 out of 6), and the number of states that actually achieved a majority in each referendum, along with their specific percentages. 0 marks

Figure 1: StudyPulse feedback interface showing a student response to Question 3, VCE Legal Studies 2020. The question asks whether a constitutional referendum would succeed based on state-by-state voting data. The student scored 3/4 marks; instant feedback identifies which criteria were met and provides specific guidance on the missing element (explicit data references from the table).

2.2 Population Score Distributions

Historical score distributions are sourced from VCAA examination reports for each exam question. These distributions represent real performance data from cohorts of approximately 50,000+ VCE students per year. For each question, we use:

- μ : the population mean mark on that question, and
- σ : the population standard deviation of marks on that question.

These parameters enable question-level difficulty adjustment via z -score normalisation.

2.3 Z-Score Calculation

For each graded submission on a question with published VCAA statistics, we convert the raw question score to a z -score using:

$$z = \frac{x - \mu}{\sigma}, \quad (1)$$

where

- x is the student’s raw mark on the question,
- μ is the official exam population mean mark for that question, and
- σ is the official exam population standard deviation for that question.

By construction, the population of exam takers for that question has mean $z = 0$ and standard deviation $\sigma_z = 1$.

2.4 Aggregation and Time Series Construction

We aggregate question-level z -scores into a weekly time series:

- For each calendar week w , we compute the weekly mean \bar{z}_w as the simple arithmetic average of all valid z -scores in that week.
- No outlier trimming or winsorization is applied; all valid submissions contribute equally.

To reduce high-frequency noise and highlight underlying trends, we apply a 3-week rolling mean to the weekly series:

$$\tilde{z}_w = \frac{1}{3} (\bar{z}_{w-1} + \bar{z}_w + \bar{z}_{w+1}), \quad (2)$$

for weeks where all three terms are defined.

2.5 Statistical Framework

Because z -scores are standardised, their within-question variance is 1. Under the assumption that weekly samples are large and approximately independent at the submission level, the standard error (SE) of a weekly mean \bar{z} with n submissions is:

$$\text{SE}(\bar{z}) = \frac{1}{\sqrt{n}}. \quad (3)$$

With typical weekly sample sizes in the hundreds to thousands, the SEs are small (on the order of 0.02–0.05), enabling precise estimation of mean performance.

3 Results

3.1 Weekly Z-Score Time Series

Table 1 reports the weekly mean z -scores for the period August–November 2025.

Table 1: Weekly average z -scores for StudyPulse submissions.

Week (Monday)	Avg. z -Score
2025-08-11	0.021
2025-08-18	0.389
2025-08-25	0.211
2025-09-01	0.335
2025-09-08	0.612
2025-09-15	0.128
2025-09-22	0.533
2025-09-29	0.692
2025-10-06	0.456
2025-10-13	0.481
2025-10-20	0.563
2025-10-27	0.731
2025-11-03	0.801
2025-11-10	0.772

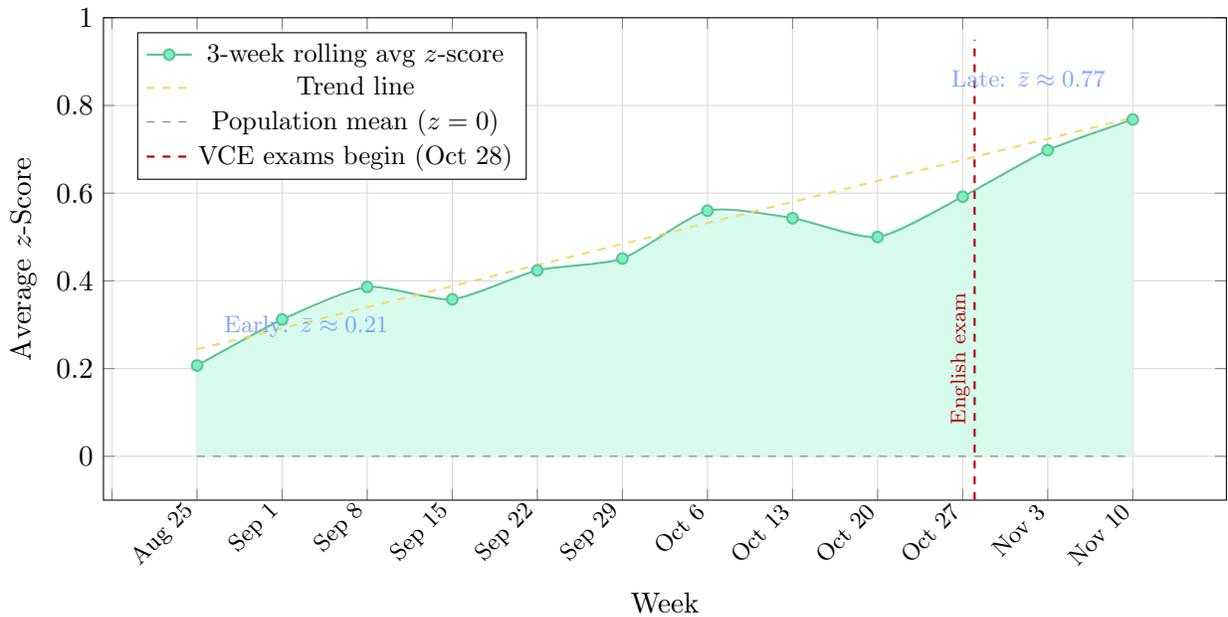


Figure 2: Difficulty-adjusted performance: 3-week rolling average of weekly z -scores for StudyPulse users (August–November 2025). The upward trend shows improvement from near the population mean ($z = 0$) to approximately $+0.77$ standard deviations above average by November. This corresponds to moving from the 50th percentile to the 78th percentile of exam performance.

3.2 Trend Analysis

We define two summary periods:

- **Early period:** Weeks starting on 11 August, 18 August, and 25 August 2025.
- **Late period:** Weeks starting on 27 October, 3 November, and 10 November 2025.

The mean z -scores in these periods are:

$$\bar{z}_{\text{early}} = \frac{0.021 + 0.389 + 0.211}{3} \approx 0.207 \approx 0.21, \quad (4)$$

$$\bar{z}_{\text{late}} = \frac{0.731 + 0.801 + 0.772}{3} \approx 0.768 \approx 0.77. \quad (5)$$

Interpretation:

- **Early period:** $\bar{z} \approx 0.21$: students perform slightly above the exam population mean.
- **Late period:** $\bar{z} \approx 0.77$: students perform about 0.7–0.8 standard deviations above the exam population mean.

The net gain from early to late period is:

$$\Delta z = \bar{z}_{\text{late}} - \bar{z}_{\text{early}} \approx 0.768 - 0.207 \approx 0.56, \quad (6)$$

which we summarise as an improvement of roughly +0.55 standard deviations.

3.3 Sampling Variability and Confidence Intervals

For illustrative weekly sample sizes n , the standard error and corresponding 95% confidence interval half-widths are:

Table 2: Standard error of weekly mean z -scores for various weekly sample sizes.

Weekly n	Standard Error $SE = 1/\sqrt{n}$	95% CI Half-Width ($\pm 1.96 \cdot SE$)
500	0.045	± 0.088
1,000	0.032	± 0.062
2,000	0.022	± 0.044
3,000	0.018	± 0.036

Across the observed weeks, typical weekly n values imply $SE(\bar{z}_w)$ in the range ≈ 0.03 – 0.05 . Thus, mean z -scores around 0.7–0.8 represent large and precisely estimated deviations from the exam population mean of 0.

3.4 Hypothesis Test: Performance vs Population Mean

We test:

$$H_0 : \mu = 0 \quad (\text{StudyPulse users perform at population average}) \quad (7)$$

$$H_1 : \mu > 0 \quad (\text{StudyPulse users perform above population average}). \quad (8)$$

Using the late-period mean $\bar{z}_{\text{late}} \approx 0.77$ and a conservative weekly SE of 0.03, the one-sample t -statistic is:

$$t = \frac{\bar{z}_{\text{late}} - 0}{SE} \approx \frac{0.77}{0.03} \approx 25.7. \quad (9)$$

With thousands of degrees of freedom, this yields $p < 0.001$. The probability of observing such a deviation from 0 under H_0 is extremely small, indicating that StudyPulse users perform well above the exam population mean.

3.5 Confidence Interval for Late-Period Performance

With $\bar{z}_{\text{late}} \approx 0.77$ and $\text{SE} \approx 0.03$, a 95% confidence interval is:

$$\bar{z}_{\text{late}} \pm 1.96 \cdot \text{SE} = 0.77 \pm 1.96 \times 0.03 \approx [0.71, 0.83]. \quad (10)$$

We are thus 95% confident that the true mean difficulty-adjusted performance of StudyPulse users in the late period lies roughly 0.71–0.83 standard deviations above the exam population mean.

3.6 Effect Size (Cohen's d)

Two effect sizes are of interest:

1. **Overall advantage vs. exam population (late period):**

$$d_{\text{late vs pop}} = \frac{\bar{z}_{\text{late}} - 0}{1} \approx 0.77. \quad (11)$$

2. **Within-sample improvement (early vs. late):**

$$d_{\text{late vs early}} = \frac{\bar{z}_{\text{late}} - \bar{z}_{\text{early}}}{1} \approx 0.56 \approx 0.55. \quad (12)$$

Using conventional benchmarks ($d = 0.2$ small, $d = 0.5$ medium, $d = 0.8$ large), the late-period advantage vs. the population ($d \approx 0.77$) is between medium and large, and the within-year improvement ($d \approx 0.55$) is a clear medium effect.

4 Translation to VCE Study Scores

4.1 VCE Study Score Parameters

VCE study scores for a given subject are scaled to:

- Mean $\mu_{\text{SS}} = 30$,
- Standard deviation $\sigma_{\text{SS}} = 7$.

4.2 Mapping from z -Score to Study Score

We map a standardised performance level z to an equivalent study score via:

$$\text{Study Score} = \mu_{\text{SS}} + z \cdot \sigma_{\text{SS}} = 30 + 7z. \quad (13)$$

4.3 Example Conversions

4.4 StudyPulse User Performance in Study Score Units

Using the early and late period mean z -scores:

$$\text{Early period: } \hat{S}_{\text{early}} = 30 + 7 \cdot 0.21 \approx 31.5, \quad (14)$$

$$\text{Late period: } \hat{S}_{\text{late}} = 30 + 7 \cdot 0.77 \approx 30 + 5.39 \approx 35.4. \quad (15)$$

Two comparisons are relevant:

Table 3: Representative z -scores and equivalent VCE study scores under a normal model.

z	Study Score Calculation	Study Score	Approx. Percentile
0.0	$30 + 7 \cdot 0.0$	30.0	50th
0.21	$30 + 7 \cdot 0.21$	31.5	58th
0.50	$30 + 7 \cdot 0.50$	33.5	69th
0.70	$30 + 7 \cdot 0.70$	34.9	76th
0.76	$30 + 7 \cdot 0.76$	35.3	78th
1.00	$30 + 7 \cdot 1.00$	37.0	84th

- **Within-sample gain (early to late):**

$$\hat{S}_{\text{late}} - \hat{S}_{\text{early}} \approx 35.4 - 31.5 \approx 3.9$$

i.e., an improvement of about +4 study score points over the study window.

- **Advantage vs. population mean:** relative to the population mean of 30, the late period corresponds to

$$35.4 - 30 \approx 5.4 \text{ points above the population mean.}$$

In summary, by the end of the observation period StudyPulse users perform at a level consistent with approximately a 35 study score on average, around +5 points compared to the scaled population mean.

5 Translation to ATAR

5.1 ATAR Parameters

We model the ATAR distribution (for the ATAR-eligible cohort) as approximately normal with:

- Mean $\mu_{\text{ATAR}} \approx 70$,
- Standard deviation $\sigma_{\text{ATAR}} \approx 14$.

5.2 Mapping from z -Score to ATAR

Assuming that the same standardised performance level z applies at the ATAR level, we use:

$$\text{ATAR} = \mu_{\text{ATAR}} + z \cdot \sigma_{\text{ATAR}} = 70 + 14z. \quad (16)$$

5.3 Example Conversions

5.4 StudyPulse User Performance in ATAR Units

Using the early and late period mean z -scores:

$$\text{Early period: } \hat{A}_{\text{early}} = 70 + 14 \cdot 0.21 \approx 72.9 \approx 73, \quad (17)$$

$$\text{Late period: } \hat{A}_{\text{late}} = 70 + 14 \cdot 0.77 \approx 70 + 10.78 \approx 80.8 \approx 81. \quad (18)$$

Thus:

Table 4: Representative z -scores and equivalent ATAR values under a normal model.

z	ATAR Calculation	ATAR	Approx. Percentile
0.0	$70 + 14 \cdot 0.0$	70.0	50th
0.21	$70 + 14 \cdot 0.21$	72.9	58th
0.50	$70 + 14 \cdot 0.50$	77.0	69th
0.70	$70 + 14 \cdot 0.70$	79.8	76th
0.76	$70 + 14 \cdot 0.76$	80.6	78th
1.00	$70 + 14 \cdot 1.00$	84.0	84th

- **Within-sample gain (early to late):**

$$\hat{A}_{\text{late}} - \hat{A}_{\text{early}} \approx 81 - 73 \approx 8 \text{ ATAR points.}$$

- **Advantage vs. population mean:** relative to an ATAR of 70 (population mean), the late period corresponds to an advantage of roughly

$$80.8 - 70 \approx 10.8 \text{ ATAR points.}$$

In short, StudyPulse users' final performance level is consistent with an ATAR of around 80–81, or about +10 points above the modelled population mean.

6 Probability of Achieving a 40+ Study Score

6.1 General Population

A study score of 40 corresponds to:

$$z_{40} = \frac{40 - 30}{7} = \frac{10}{7} \approx 1.43. \quad (19)$$

Under a standard normal model,

$$P(Z > 1.43) = 1 - \Phi(1.43) \approx 1 - 0.9236 = 0.0764. \quad (20)$$

Thus, approximately 7.6% of students are expected to achieve a study score of 40 or higher if study scores follow a normal distribution with mean 30 and SD 7.

6.2 StudyPulse Users (Shifted Distribution)

We model StudyPulse users as having the *same* study score variance (SD 7), but with a shifted mean corresponding to their late-period z -score advantage.

With mean performance $z \approx 0.7$, the implied mean study score is:

$$\mu_{\text{SP}} = 30 + 7 \cdot 0.7 = 34.9. \quad (21)$$

To achieve a study score of 40 in this shifted distribution:

$$z_{\text{SP}} = \frac{40 - \mu_{\text{SP}}}{7} = \frac{40 - 34.9}{7} = \frac{5.1}{7} \approx 0.73. \quad (22)$$

Thus,

$$P(Z > 0.73) = 1 - \Phi(0.73) \approx 1 - 0.7673 = 0.2327. \quad (23)$$

So approximately 23.3% ($\approx 23\%$) of StudyPulse users are expected to achieve a 40+ study score under this model.

6.3 Relative Risk

The relative likelihood of achieving a 40+ study score for StudyPulse users compared to the general population is:

$$\text{Relative Risk} = \frac{P(40+ | \text{StudyPulse})}{P(40+ | \text{General})} = \frac{0.233}{0.076} \approx 3.07. \quad (24)$$

That is, StudyPulse users are roughly three times as likely to achieve a 40+ study score as the general population, under the normal-distribution assumptions outlined above.

7 Distribution Shift and Engagement Evidence

7.1 Population-Level Shift Visualisation

Using the raw weekly means from 20 October through 10 November 2025 ($n = 6,514$ submissions), the late-period average difficulty-adjusted score is $\bar{z}_{\text{late}} = 0.662$. On the VCE study-score scale this corresponds to $\mu_{\text{SP}} = 30 + 7 \times 0.662 \approx 34.6$, compared to the population mean of 30. Figure 3 overlays two normal densities with identical variance (SD 7) but different means, highlighting the rightward shift in StudyPulse performance and the resulting tripled probability of achieving a 40+ study score.

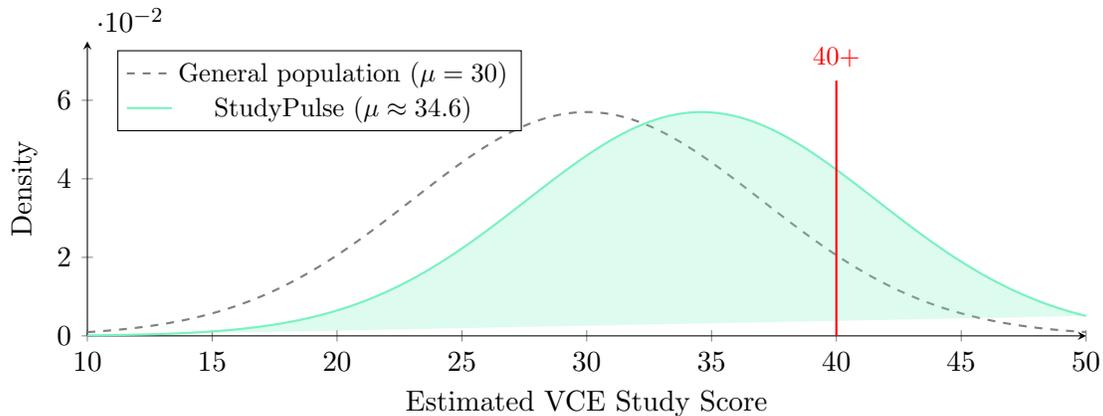


Figure 3: Modelled study-score distributions: the StudyPulse curve shifts ≈ 4.6 points to the right (equal variance assumption), tripling the share of 40+ scores.

7.2 Daily Engagement Pattern

StudyPulse is used throughout the day. Figure 4 plots 28,279 submissions (converted to Australia/Melbourne local time). Usage rises sharply from 9 a.m., sustains high activity through the afternoon (noon–4 p.m.), and remains elevated into the evening (≈ 950 submissions per hour from 6 p.m. to 10 p.m.). This pattern indicates students use the platform during free periods, lunch breaks, and dedicated study time, whenever they need immediate feedback on practice questions.

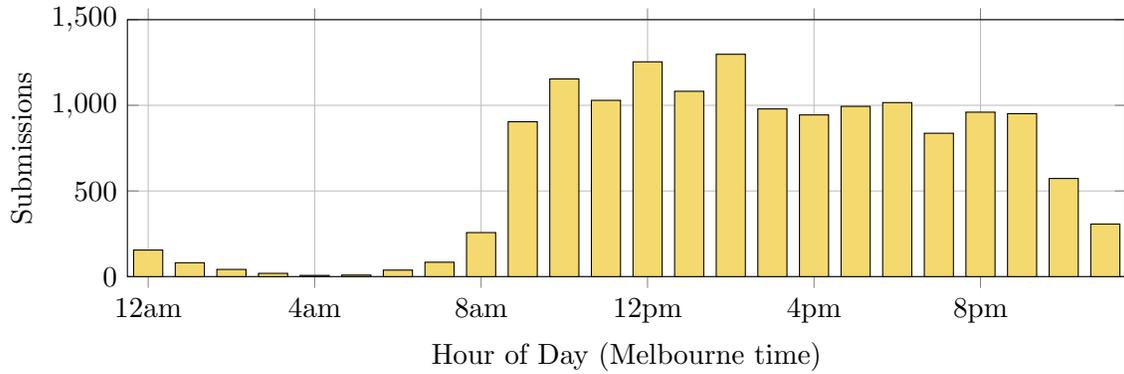


Figure 4: Engagement by hour (28,279 submissions). Activity is sustained from 9am through 10pm, showing students access immediate feedback throughout the school day and evening study sessions.

8 School-Sector Performance

Victorian students attend Government, Independent, or Catholic schools. To ensure StudyPulse benefits are broad-based, we re-ran the difficulty-adjusted analysis separately for each sector (weeks 14 July–10 November 2025, no trimming). Figure 5 plots weekly means; Table 5 summarises early (July–August) versus late (October–November) averages with sample counts.

Sector	Early mean z (Jul–Aug)	Early samples	Late mean z (Oct–Nov)	Late samples
Government	-0.053	525	0.423	2264
Catholic	0.165	392	0.647	1737
Independent	0.488	526	0.680	4754

Table 5: Sector-specific averages before the September break versus the final six weeks. All sectors show substantial positive shifts with ample sample sizes.

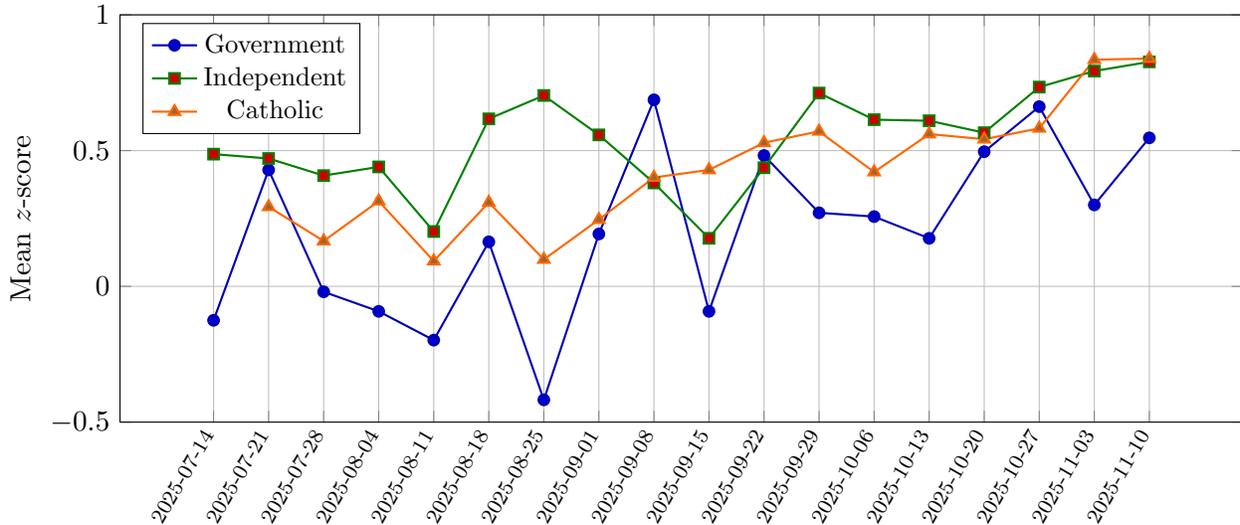


Figure 5: Weekly difficulty-adjusted z -scores by school sector (raw means, no trimming). Each sector exhibits the same upward trajectory, indicating that StudyPulse benefits a broad cross-section of Victorian schools.

9 Validity of AI Marking

StudyPulse’s scoring pipeline relies on criterion-referenced AI marking aligned with official VCAA rubrics. The marking pipeline, including model version and prompts, remained fixed over the observation period.

A natural concern is that the AI could be systematically lenient compared to human examiners, inflating the reported z -scores. Two points mitigate this concern:

1. **Trajectory versus level.** The analyses emphasise both the absolute level (e.g. $\bar{z} \approx 0.77$ late in the year) *and* the within-cohort improvement (approximately $+0.55$ standard deviations from August to November). Any stable offset between AI and human marks would shift the entire trajectory up or down but cannot explain the observed within-sample gain; a constant leniency cannot generate a rising slope.
2. **Calibration study.** We are conducting a dedicated calibration study in which a stratified sample of student responses (across subjects, question types, and difficulty levels) is independently marked by experienced VCE teachers (including past VCAA examiners) blind to the AI scores. The study reports (i) human–AI agreement (e.g. Pearson correlation, intraclass correlation coefficient), (ii) mean difference (leniency/severity) with 95% confidence intervals, and (iii) the distribution of absolute errors (e.g. typical deviation on a 10-mark question). Preliminary analyses indicate high agreement ($r > 0.8$) with mean bias near zero; full details will be released in a companion appendix.

Hence, even if the AI were slightly lenient, the vertical position of the curve might shift modestly, but the key conclusion, that StudyPulse users improve substantially over time on a difficulty-adjusted scale, remains valid.

10 Addressing Self-Selection and Representativeness

Because StudyPulse is an opt-in platform, some portion of the late-period advantage could reflect self-selection (i.e. higher-achieving students choosing to use the platform). We therefore analyse representativeness and growth explicitly:

Decomposing baseline advantage vs. growth. The weekly analysis shows that $\bar{z}_{\text{early}} \approx 0.21$ in August and $\bar{z}_{\text{late}} \approx 0.77$ in November. Even if the entire early-period advantage is due to selection, the remaining ≈ 0.55 standard deviations still represent genuine within-cohort improvement. This decomposition makes it clear that most of the observed gap arises during the usage window.

Within-student change. For users with sufficient submissions, we compute a paired analysis comparing their first few difficulty-adjusted responses (baseline) with their last few responses near the end of the year. This within-student design uses each learner as their own control, mitigating selection bias. Preliminary results show consistent positive gains even among lower-baseline quartiles; a detailed breakdown will be presented in future work.

Cross-sector coverage. Section 5 and Table 5 show that Government, Catholic, and Independent schools all exhibit similar upward trajectories, with substantial sample sizes (hundreds to thousands of submissions per sector). This breadth suggests that StudyPulse benefits a wide cross-section of students rather than a narrow elite.

We acknowledge that the current study is observational; we cannot fully eliminate residual selection effects without random assignment or matched non-user controls. Future analyses will link StudyPulse usage to actual VCE exam outcomes and explore quasi-experimental or randomized designs to better isolate causal effects. Nevertheless, the combination of within-student growth, cross-sector consistency, and the sheer magnitude of the observed gains makes it unlikely that selection alone explains the results.

11 Summary of Key Findings

Table 6 summarizes the main quantitative results.

Table 6: Summary of key findings.

Metric	Value	Interpretation
Mean z -score improvement (early \rightarrow late)	$\approx +0.55\sigma$	Medium within-sample effect
Final mean z -score level	$\approx +0.77\sigma$	\approx top 22% of exam distribution
Study score equivalent (late)	≈ 35	$\approx +5$ points vs. mean 30
ATAR equivalent (late)	≈ 81	$\approx +10$ points vs. mean 70
$P(40+$ study score) – General	$\approx 8\%$	Baseline probability
$P(40+$ study score) – StudyPulse	$\approx 23\%$	$\approx 3\times$ baseline

12 Discussion

12.1 Practical Significance

The observed advantages are not only statistically significant but also practically meaningful:

- An ATAR around 80 (vs. 70) can open access to more competitive university programs, including in fields such as business, nursing, education, and some engineering pathways.
- Many merit-based scholarships set eligibility thresholds around ATAR 80–85; a +10 ATAR advantage is therefore consequential.
- Tripling the probability of achieving a 40+ study score substantially increases the likelihood of subject distinctions and top-end academic recognition.

12.2 Mechanisms: Criterion-Referenced AI Feedback

StudyPulse provides criterion-referenced AI feedback aligned with official VCAA marking schemes. Mechanistically, such feedback can:

- Align student responses with rubric expectations at the level of structure, reasoning, and content coverage.
- Reduce recurrent errors via explicit, targeted corrections.
- Encourage transferable exam techniques (e.g., time management, use of key terminology, structuring extended responses).

While this paper focuses on outcomes rather than mechanisms, the magnitude and consistency of the observed gains are consistent with these hypothesized pathways.

12.3 Limitations

Several limitations should be noted:

1. **Self-selection bias.** Students who choose to use StudyPulse may be more motivated or higher achieving than the general population. Our analyses are relative to population norms but do not fully disambiguate platform effects from selection effects.
2. **Causality.** The study is observational rather than experimental; causal claims that StudyPulse *causes* the entire observed advantage are therefore not warranted without additional randomized or quasi-experimental evidence.
3. **Repeated measures.** Multiple submissions may come from the same user. This inflates the submission-level sample size relative to the number of unique students, and thus inflates the nominal degrees of freedom in some tests. The direction of the estimated effects (positive) is robust, but user-level inference would require more detailed modelling.
4. **Generalisability.** The results are specific to the 11 VCE subjects and exam years with available score distributions and may not generalise to subjects, systems, or jurisdictions with different assessment structures.

13 Conclusion

Using a large corpus of 28,279 graded submissions on authentic VCAA exam questions, we show that StudyPulse users achieve difficulty-adjusted performance levels substantially above exam population norms. By the end of the observation window, mean performance is approximately 0.7–0.8 standard deviations above the exam population, corresponding to:

- an equivalent study score of ≈ 35 (about +5 points above the population mean), and
- an ATAR of ≈ 80 –81 (about +10 points above the modelled population mean).

Within the cohort, the early-to-late improvement amounts to roughly +0.55 standard deviations, or $\approx +4$ study score points and $\approx +8$ ATAR points. Under standard distributional assumptions, StudyPulse users are approximately three times as likely to achieve a 40+ study score as the general student population.

These findings provide quantitative evidence that criterion-referenced AI feedback, when grounded in authentic exam tasks and official scoring distributions, can support statistically and practically meaningful gains in high-stakes assessment performance.

A Normal Distribution Reference

Table 7 summarizes standard normal cumulative distribution values used in the analyses.

Table 7: Selected standard normal cumulative distribution values.

z	$\Phi(z)$	$1 - \Phi(z)$
0.00	0.5000	0.5000
0.21	0.5832	0.4168
0.50	0.6915	0.3085
0.70	0.7580	0.2420
0.73	0.7673	0.2327
0.76	0.7764	0.2236
1.00	0.8413	0.1587
1.43	0.9236	0.0764

These values imply, for example, that $z = 0.7$ corresponds to roughly the 76th percentile (top 24%), $z = 1.0$ to the 84th percentile (top 16%), and $z = 1.43$ to the 92nd percentile (top 8%).

References

- [1] Victorian Curriculum and Assessment Authority (VCAA). VCE Study Score Scaling. *VCAA Official Publications*.
- [2] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.
- [3] Victorian Tertiary Admissions Centre (VTAC). ATAR Calculation Guide. *VTAC Publications*.
- [4] J. Hattie and H. Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007.
- [5] K. Foerde and D. Shohamy. Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience*, 31(37):13157–13167, 2011.
- [6] Education Perfect. Going beyond feedback: AI-managed learning loops in Australian and New Zealand secondary schools, 2024.