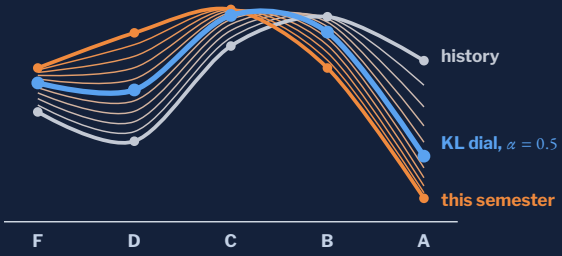
What does an “optimal” grade distribution look like?

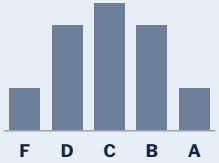


WARM-UP

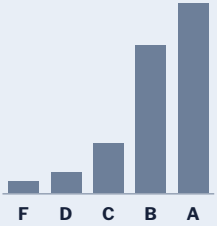
Sketch the distribution you would defend



Flat



Bell



Mostly A and B



Bimodal

Which shape would you defend to a colleague? To a student? To a dean?

What curve do you implement?

Example (hard final)

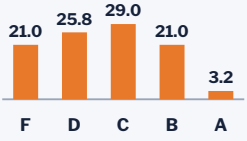
- **124 students**
- **46.8%** below C (typical semester has **26%**)

Do you curve? By how much? What would make the answer principled rather than a feeling?

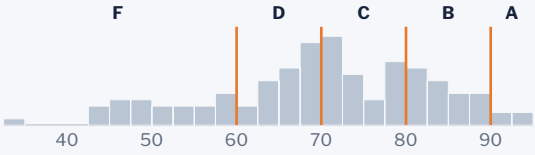
Typical semester
history, the prior q · 26% below C



This semester: hard final
no curve · 46.8% below C



Hard final scores
students per 2.5 points; orange lines are the no-curve cutoffs 60/70/80/90



History is an example department prior: A 22 · B 28 · C 24 · D 11 · F 15.

1 · THE QUESTION

Distributions decide what grades can say

43%

of US college letter grades were A's by 2009. When most grades match, each says less.

Rojstaczer & Healy, 2012

11%

less information in Princeton grades than thirty years earlier, measured in bits.

Felten, 2004

27%

of Calc I students nationally get D, F or W. In a gateway course, the bottom of the distribution decides who continues.

MAA CUPM calculus report, from CSPCC data

Features of grades

Feedback

Tells the student where they stand and what to work on next.

Certification

Says the student is ready for the next course.

Signal

Tells advisors, the next instructor, and employers what they cannot see directly.

Incentive

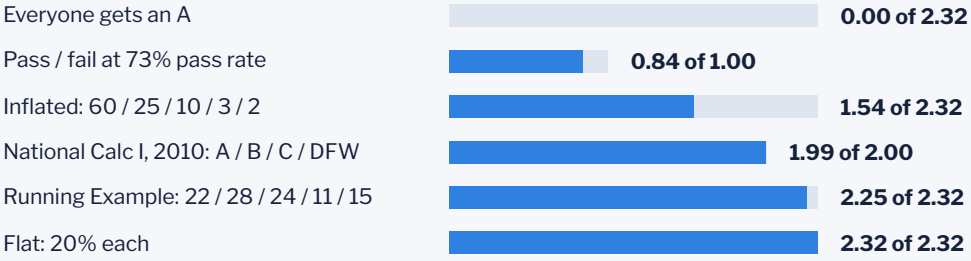
Shapes effort now, and who decides to stay in STEM.

Which job does your current grading scheme do best? Which does it do worst?

Entropy measures how much one grade tells you

$$H(p) = -\sum_{g} p(g)\log_2 p(g)$$

Bits: the average number of yes-or-no questions needed to learn a student's grade. With k grades the most it can be is $\log_2 k$.



The national row is from a fall 2010 MAA CSPCC survey.

2 · GRADES AS SIGNALS

Maximize information, and every grade should be equally common.

For five letter grades that means 20% each and $\log_2 5 \approx 2.32$ bits per grade. Almost nobody would aim for that in a course like Calculus I.

KL divergence

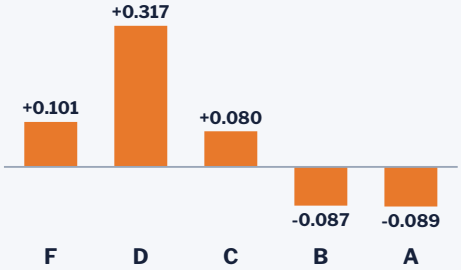
$$D(p \,\|\, q) = \sum_g p(g)\, \log_2 \frac{p(g)}{q(g)}$$

q is your prior
The course's history, what a typical term looks like.

$D \geq 0$, and $D = 0$ only when $p = q$.
Measured in bits, it is the extra cost of reading this semester's grades with history's expectations.

Maximum entropy is a special case.
$D(p \,\|\, \text{uniform}) = \log_2 k - H(p)$, so with a flat prior, minimizing KL is maximizing entropy.

Where the hard final's 0.322 bits come from



Each bar is one term $p(g)\log_2[p(g)/q(g)]$ for the hard final with no curve: below zero where a grade is rarer than history, above where it is more common. They add to 0.322; the D's alone give 0.317.

Why KL divergence?

If a class of N students were a random draw from history q, the chance its grades look like p is about $2^{-N \cdot D(p\,\|\,q)}$ (Sanov's theorem).

$$2^{-N \cdot D(p\,\|\,q)}$$

The chance band

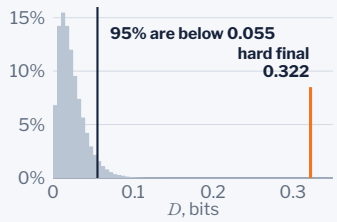
How far a class of N lands from history by chance alone: **95% of random classes stay closer.**

Simulate: draw N grades from q, compute D, repeat (middle).

Formula: $2N \ln 2 \cdot D$ is about χ^2_{k-1}, so with $k = 5$ grades $D_{95} \approx 9.49/(2N \ln 2) \approx 6.8/N$ bits (right).

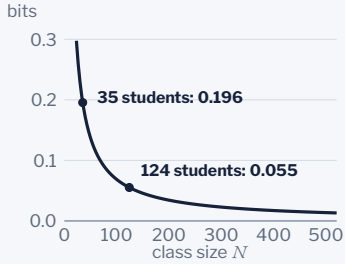
Chance alone: classes of 124

D of 200,000 random classes from q



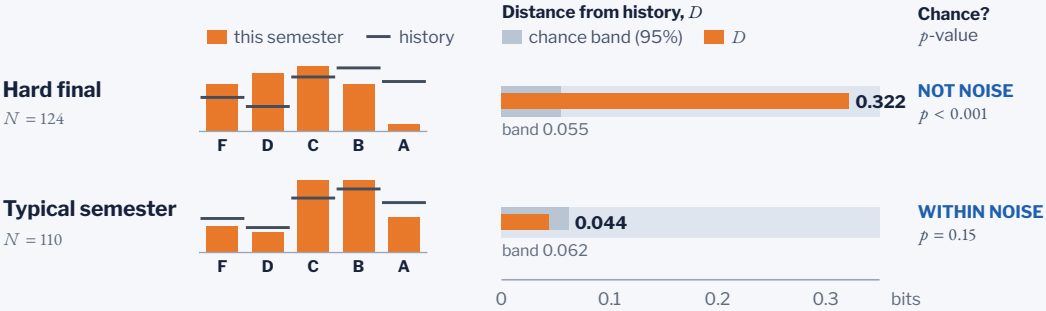
The 95% band, for every class size

$D_{95} \approx 6.8/N$ bits (formula)

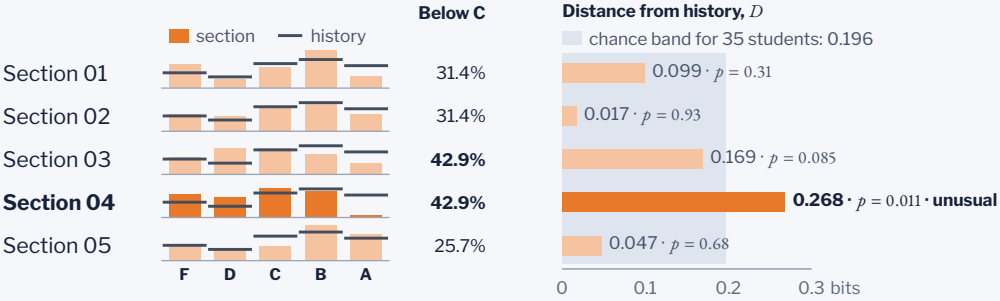


For 35 students, simulation gives 0.206 rather than the formula's 0.196.

Is this semester unusual?



Checking across sections



What is the right conversation to have with Section 4's instructor, and what evidence would you want first?

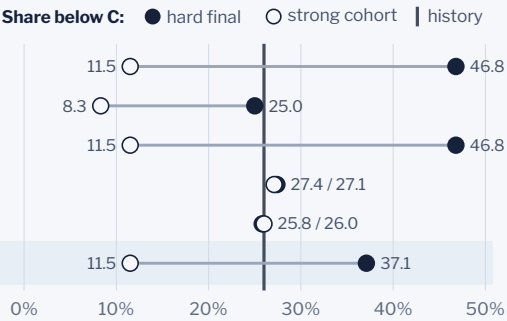
Scores alone cannot tell a hard exam from a weaker cohort

	Below history like the hard final: 46.8% below C	**Above history** like a strong cohort: 11.5% below C
		
The exam was off too hard or too easy ■ curve to history leans here	Curve up	Curve down
The cohort really differs weaker or stronger ■ no curve leans here	Leave the grades	Leave the grades

For the running example, we know the truth: the hard final is a typical class minus 7 points.

Common curving rules

Rule	How it works
No curve	90/80/70/60 on raw scores
Top score raised to 100	same points to all: +7.0 / +1.9
Median at least 70	adds points only if median < 70
Bell curve	grades by z-score bands
Curve to history	cutoffs by rank to match history
KL dial, $\alpha = 0.5$	never below no curve



Running examples: hard final $N = 124$ (median 70.8, top 93.0); strong cohort $N = 96$ (median 86.2, top 98.1).

Weighing history against this semester

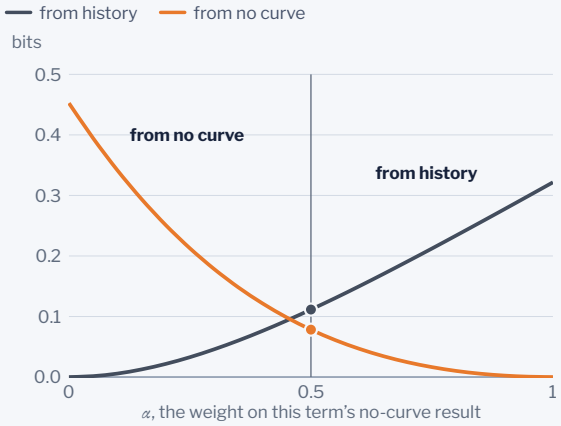
$$p_\alpha(g) \propto q(g)^{1-\alpha}\, s(g)^{\alpha}$$

s is this semester's grade distribution with no curve. The target grade distribution is the unique minimizer of

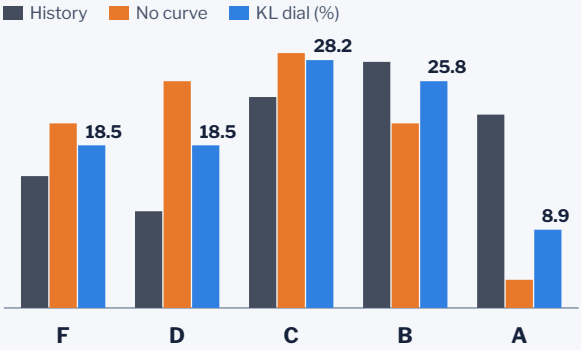
$$(1-\alpha)\, D(p \,\|\, q) + \alpha\, D(p \,\|\, s)$$

At $\alpha = 0.5$ on the hard final, the target sits 0.111 bits from history and 0.078 bits from no curve.

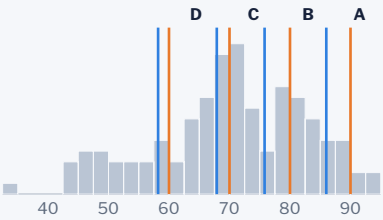
α is the weight you put on "the cohort really differs."
$\alpha = 1$ is no curve, $\alpha = 0$ is curve to history



The KL divergence dial at $\alpha = 0.5$



The cutoffs move down a few points



Cutoff	D	C	B	A
No curve	60	70	80	90
KL dial	58.2	67.9	75.8	86.0

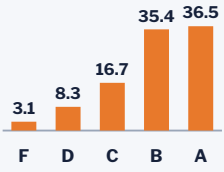
Below C: 46.8% → 37.1%. 35 students move up a grade; none move down.

On a strong cohort, the same rule needs a guard

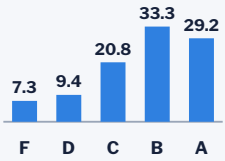
History
the prior $q \cdot 26\%$ below C



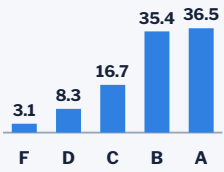
Strong cohort, no curve
$N = 96 \cdot 11.5\%$ below C



KL dial, $\alpha = 0.5$, no guard
25 students move down a grade



KL dial, $\alpha = 0.5$, guarded
nobody moves: same as no curve



The guard

No student gets a lower grade than with no curve.

Is a rule that only ever moves grades up fair to the next cohort, and to what a transcript means?

Without the guard the dial pulls this cohort toward history; with it, cutoffs only move down.

6 · LIMITS

What KL divergence cannot tell you

Whether the scores are valid

KL divergence measures distributions, not learning. A badly written exam can produce a confident yet wrong signal.

Whether the prior is fair

History encodes past placement, past grading cultures and past inequities. Matching it can preserve them.

Why a semester differs

Hard exam or weaker cohort? Common anchor items (e.g. uniform quiz) reused across terms can answer what the grade distribution cannot.

What rank-based cutoffs do

Research ties curved grading in intro STEM to a competitive climate that falls hardest on underrepresented students.

See: EAB on norm-referenced grading; HERI (ASHE 2014); ASEE on left-of-center grading; Seymour & Hunter (2019).

7 · DEMO ·

Hands-on with the KL Divergence

1 · Turn the dial

Load “hard final.” Find the α where Below C first drops under 40%. Would you publish that α in a syllabus?

2 · Which bet pays off?

In step 5, try “Only the exam varies,” then “Only the cohort varies.” Which rule wins each time? Where does the best α land?

Later · Bring your own

Upload past terms (scores, or registrar grade counts) and this term’s scores. Is your term within noise?

Demo link: seangrate.com/kl/

8 · DISCUSSION ·

Questions for the room

- What should our prior q be: the last five years, national data, or a target we choose on purpose?
- Who sets α, and is it set before or after the scores are in?
- Does a “never lower” guard make grades more honest or only more inflated?

What should a grade distribution look like?

1

Informative.

Grades are messages. Bits are a currency to measure information, so inflation and compression of grades changes the value of the grades.

2

Close to history, unless the evidence says otherwise.

KL divergence measures a semester's departure from the expectation.

3

Every curving scheme bets between history and this semester's scores.

Use a parameter to balance between historic trends and this semester's features.

SOURCES

Sources and further reading

- Academia Stack Exchange, “What does an optimal grading distribution look like?” (question 61038, answer 61040).
- Bressoud & Rasmussen, “Seven characteristics of successful calculus programs,” *Notices of the AMS* 62(2), 2015; MAA CSPCC national survey, fall 2010.
- MAA CUPM Curriculum Guide 2015, Calculus course report.
- Rojstaczer & Healy, “Where A is ordinary,” *Teachers College Record*, 2012. Felten, “How much information do Princeton grades convey?” CITP blog, 2004.
- Chan, Li & Suen, “A signaling theory of grade inflation,” *International Economic Review*, 2007. Dubey & Geanakoplos, “Grading exams: 100, 99, 98, … or A, B, C?” *Games and Economic Behavior*, 2010.
- Micha, Sekar & Shah, “What is best for students, numerical scores or letter grades?” IJCAI 2024.
- Kullback, *Information Theory and Statistics*, 1959. Cover & Thomas, *Elements of Information Theory* (Sanov’s theorem).
- Norton, Bridges & High, “Impact of course policy changes on Calculus I DFW rates,” *Journal of STEM Education* 19(1), 2018.
- Seymour & Hunter (eds.), *Talking about Leaving Revisited*, 2019. HERI, “Competition in introductory STEM courses,” ASHE 2014.