

What's old is new:
Tutorial-style homework and improved learning

John Duchi
(joint with William Hartog and Dennis Sun)
Stanford University

October 2026

We talkin' 'bout practice¹

why we like(d) homework:

- ▶ measures learning
- ▶ elicits effort outside of class
- ▶ induce productive struggle and persistence

We talkin' 'bout practice¹

why we like(d) homework:

- ▶ measures learning
- ▶ elicits effort outside of class
- ▶ induce productive struggle and persistence

educational psychology [Dunlosky et al., 2013]: two super effective learning strategies include

- ▶ *practice testing*
- ▶ *distributed practice* (i.e., not cramming for the test)

weekly problem sets implement both

¹AI

LLMs (probably) hurt learning

- ▶ high-school math: AI without guardrails impedes learning [[Bastani et al., 2025](#)]
- ▶ 50,000 Chinese high-school students over 4 years [[Strömberg et al., 2026](#)]: with AI,
 - ▶ better, faster homework
 - ▶ worse exam performance
- ▶ lower-quality reasoning [[Stadler et al., 2024](#)], reduced persistence and worse unassisted performance [[Liu et al., 2026](#)]
- ▶ brain activity lower [[Kosmyna et al., 2025](#)]

The goal

Incentivize doing homework to maintain practice
(and maybe some other stuff as well)

The goal

Incentivize doing homework to maintain practice
(and maybe some other stuff as well)

- ▶ abandoning homework means students just cram
- ▶ in-class exams measure learning, pacing incentive unclear
- ▶ good luck policing AI use

An old answer: the Oxbridge tutorial

- ▶ Oxford “tutorials” / Cambridge “supervisions”: small groups (1–6 students) meet regularly with tutor [[Ashwin, 2005](#)]
- ▶ students work problems beforehand, tutor leads discussion of their answers
- ▶ tutoring associated with large gains in learning [[VanLehn, 2011](#)]

An old answer: the Oxbridge tutorial

- ▶ Oxford “tutorials” / Cambridge “supervisions”: small groups (1–6 students) meet regularly with tutor [[Ashwin, 2005](#)]
- ▶ students work problems beforehand, tutor leads discussion of their answers
- ▶ tutoring associated with large gains in learning [[VanLehn, 2011](#)]

our adaptation:

- ▶ same problem sets, but written part ungraded
- ▶ weekly oral: randomized **variant** problem in a group of four

Setting: Stats 117, Spring 2026

- ▶ Introduction to Probability: mostly first- and second-year undergraduates
- ▶ 10 weeks, 29 lectures, one 6ish-problem set per week
- ▶ Counting, discrete random variables, expectation and variance, conditioning, continuous distributions

Treatment (tutorials)

- ▶ hand in homework, ungraded
- ▶ weekly 50-min tutorial replaces homework grade

Control (standard)

- ▶ homework graded as usual
- ▶ weekly 50-min TA problem-solving section (to match time)

- ▶ same lectures, same outcome measure: proctored midterm and final

A tutorial session

- ▶ 3–4 students, 1 tutor (TA or professor), 50 minutes
- ▶ each student presents once: 12 minutes at the board, no notes
- ▶ problem is *variant* of homework problem
 - ▶ revealed at student's turn
 - ▶ presenter selected in random order
- ▶ others: stay off devices, follow argument, help when presenter is stuck

A tutorial session

- ▶ 3–4 students, 1 tutor (TA or professor), 50 minutes
- ▶ each student presents once: 12 minutes at the board, no notes
- ▶ problem is *variant* of homework problem
 - ▶ revealed at student's turn
 - ▶ presenter selected in random order
- ▶ others: stay off devices, follow argument, help when presenter is stuck

grading: score $\{0, 1, 2\}$ for effort and engagement, not correctness

- ▶ exams to measure correctness
- ▶ getting stuck on hard variant is safe(ish)

Variant problems

idea: limit bog-standard memorization/staring at laptop in session

Example: Counting questions

HW: In Texas Hold'em (2 in hand + 5 community cards), what is the probability of a flush (i.e., that at least 5 of the 7 cards share a suit)?

Tutorial: In Bridge (13 cards each in partnerships of two of four individuals), what is the probability your partnership is long, holding ≥ 10 cards of some suit?

Randomization

- ▶ homework 0: 12-question concept-inventory pretest
 - ▶ independence, geometric sampling, with/without replacement, expectation, Bayes' rule
- ▶ **stratification**: within each pretest score, randomly split students between treatment (tutorials) and control (sections)

Pretest score	Tutorials	Control
12/12	3 students	2 students
9/12	5 students	5 students
⋮	⋮	⋮

- ▶ 80 students in analysis: 38 tutorials, 42 control (77 took final)
- ▶ IRB #83954

Main results

Students in tutorials scored

≈ 15 points higher on the midterm (of 101)

≈ 18 points higher on the final (of 154)

$n = 66$ (of 80 in class; results consistent with full class),

$$p \leq .002$$

Main results

Students in tutorials scored

≈ 15 points higher on the midterm (of 101)

≈ 18 points higher on the final (of 154)

$n = 66$ (of 80 in class; results consistent with full class),

$$p \leq .002$$

- randomized, stratified on a pretest, same exams

Formal tests

randomization test (null = no individual gains)

n	Outcome	Difference	Null SD	p -value
66	Midterm	15	5.0	.002
64	Final	15	5.8	.002

standard factor analysis for average treatment effect

$$\tau = \mathbb{E}[Y(\text{treat})] - \mathbb{E}[Y(\text{control})]$$

n	Outcome	$\hat{\tau}$	std. error	p -value	95% lower bound
66	Midterm	15.1	4.9	.002	6.9
64	Final	17.6	5.6	.001	8.2

units of exam points (midterm out of 101, final out of 154)

Caveats

- ▶ **time on task:** did not enforce section attendance
- ▶ **mechanism:** cannot separate oral assessment, small-group contact, peer effects, preparing to teach, ...
- ▶ **scope:** one course, one quarter, one John, one Stanford
- ▶ **permission:** 66 of 80 students permitted inclusion of their data, though results on that subset consistent

Caveats

- ▶ **time on task:** did not enforce section attendance
- ▶ **mechanism:** cannot separate oral assessment, small-group contact, peer effects, preparing to teach, ...
- ▶ **scope:** one course, one quarter, one John, one Stanford
- ▶ **permission:** 66 of 80 students permitted inclusion of their data, though results on that subset consistent

now:

- ▶ larger study running across 5 classes
- ▶ different interventions

Costs

- ▶ typical staffing: one TA per 30–50 students
- ▶ 30 students in groups of 4 means 7–8 tutorial hours/week
 - ▶ reasonable(?) for a 50% (20 hour/week) TAship
 - ▶ eliminates homework grading
- ▶ need variant problems
- ▶ idea: get undergraduate tutors
- ▶ win-win: teaching means tutors learn

Costs

- ▶ typical staffing: one TA per 30–50 students
- ▶ 30 students in groups of 4 means 7–8 tutorial hours/week
 - ▶ reasonable(?) for a 50% (20 hour/week) TAship
 - ▶ eliminates homework grading
- ▶ need variant problems
- ▶ idea: get undergraduate tutors
- ▶ win-win: teaching means tutors learn

Open questions:

- ▶ do 1-on-1 or shorter sessions work as well?
- ▶ remote or larger courses?
- ▶ carry over to other courses?

Take-homes

- ▶ homework remains a good idea
- ▶ 4Cs: **Communication**, Creativity, Collaboration, **Critical Thinking**
- ▶ in randomized trial, tutorials raised exam scores by ≈ 15 points, robustly
- ▶ cost: TA hours we nominally already spend²

²but I'd find leading 7 hours of tutorials mind-numbing

Analysis: potential outcomes

For student i , exam score $Y_i(1)$ with tutorials, $Y_i(0)$ without;
treatment indicator Z_i

1. Randomization test [Caughey et al., 2023]

$$H_{0,c} : Y_i(1) \leq Y_i(0) + c \text{ for each } i$$

- ▶ re-randomize Z only *within* pretest strata (10^6 draws);
statistic is the difference in means
- ▶ report p -value at $c = 0$ and largest rejected $\hat{c}$

Analysis: potential outcomes

For student i , exam score $Y_i(1)$ with tutorials, $Y_i(0)$ without; treatment indicator Z_i

1. Randomization test [Caughey et al., 2023]

$$H_{0,c} : Y_i(1) \leq Y_i(0) + c \text{ for each } i$$

- ▶ re-randomize Z only *within* pretest strata (10^6 draws); statistic is the difference in means
- ▶ report p -value at $c = 0$ and largest rejected $\hat{c}$

2. Stratified Neyman z -test [Imbens and Rubin, 2015]

$$\tau = \mathbb{E}[Y(1) - Y(0)], \quad \hat{\tau} = \sum_k \frac{n_k}{n} (\bar{Y}_{1k} - \bar{Y}_{0k})$$

- ▶ merge adjacent pretest strata so each has ≥ 2 per arm
- ▶ conservative variance; one-sided $H_0 : \tau \leq 0$

Bibliography I

- P. Ashwin. Variation in students' experiences of the "Oxford tutorial". *Higher Education*, 50(4):631–644, 2005.
- H. Bastani, O. Bastani, A. Sungu, H. Ge, Özge Kabakcı, and R. Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025.
- D. Caughey, A. Dafoe, X. Li, and L. Miratrix. Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. *Journal of the Royal Statistical Society, Series B*, 85:1471–1491, 2023.
- J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.
- G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Bibliography II

- N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv:2506.08872 [cs.AI]*, 2025.
- G. Liu, B. Christian, T. Dumbalska, M. A. Bakker, and R. Dubey. AI assistance reduces persistence and hurts independent performance. *arXiv:2604.04721 [cs.AI]*, 2026.
- M. Stadler, M. Bannert, and M. Sailer. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160(108386):1–7, 2024.
- D. Strömberg, V. Lei, and Y. Wu. The generative AI learning penalty: Evidence from Chinese secondary education. *Centre for Economic Policy Research*, (DP21577), 2026. Discussion Paper.
- K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.