

When AI Closes the Task Gap

What experiments reveal, and leave unresolved, about the value of a qualification.

Imagine being asked to answer your manager's email about a business problem. The evidence is in a short report: text, a chart and a table. You must diagnose what went wrong and propose a solution. In an experiment in Argentina, adults with more education scored substantially higher than adults with less education when neither group was assigned an AI assistant. Give both groups access to one, and the difference became much smaller. The measured gap fell from **0.548 to 0.139 standard deviations**, about three quarters of the initial gap on that task.[1]

The result tempts a leap from task scores to the value of a qualification. But the experiment observed neither hiring nor pay. A second experiment, among college students, tested knowledge after AI was removed. It found higher scores about a week later, although compliance concerns limit what those scores establish about unaided learning.[2] Historical wage research points to another route connecting education and earnings: the jobs people enter and the scope those jobs offer for later growth.[3]

These studies separate three possible routes from AI to earnings: assisted output today, learning that lasts beyond the task, and access to work that pays. The evidence is strongest for the first, narrower for the second, and does not settle the third.

A smaller gap on the same task

The Argentine researchers recruited people aged 25–45 between September and November 2025 and randomly assigned access to an AI assistant for an online task. Participants answered a hypothetical manager's email using supplied information; the problem concerned a cafeteria, food-delivery service or appliance store. Everyone had what they needed to write a diagnosis and solution. The design isolates assigned assistance on a self-contained problem, while leaving a firm's accumulated knowledge, relationships and continuing work outside the test.[1]

The education categories need care. The study's "lower-education" group had a high-school diploma and either no postsecondary study or less than half of a university or non-university tertiary programme completed. Its "higher-education" group had completed more than half of such a programme, including people who had **not** finished a qualification. These are pre-existing groups, not people randomly assigned to complete a degree. The experiment tests whether *AI access has different effects across those groups*; it cannot identify what gaining a qualification would have done to any participant.[1]

Among the 1,174 completers, the higher-education group led by 0.548 standard deviations without assigned AI access and 0.139 with it. The estimated change in the gap was **−0.408 standard deviations** (standard error 0.122; approximate 95% interval −0.647 to −0.169). Scores use the lower-education control group's standard deviation. Subtracting the printed gaps gives −0.409 because of rounding. The remaining gap is less certain than the compression: its approximate interval runs from −0.032 to 0.310. The result establishes neither equality nor a persisting advantage.[1][4]

The compressed difference was in *scored output* on this problem. A rubric combined diagnosis, proposed solution and writing quality. The researchers averaged ten AI grading iterations, checked 117 randomly selected responses with two human graders blind to education and treatment, and ran alternative grading checks. Those checks make a simple scoring artifact less plausible. They do not turn one answer into evidence of reliability across a working week or an employer's willingness to hire its author.[1]

Completion and control-group behaviour narrow the interpretation. Of 1,795 eligible people randomized, 1,174 finished: 520 in the lower-education group and 654 in the higher-education group. Completion was 64% with assigned AI access and 66% in control. Completers looked similar on observed characteristics, but unobserved selection remains possible. The authors estimate that roughly 13% of controls may have sought AI help. The comparison is therefore assigned access versus control, with no certain direction of bias for the education-group contrast.[1]

Within those boundaries, assigned AI access sharply compressed a pre-existing score gap on a task where everyone received the same information. Many jobs demand information and judgment that cannot be supplied in a short brief.

What is left when the assistant leaves?

Immediately afterwards, all participants answered related questions without the assistant. They explained the root cause and answered interpretation and recall questions about the same material. This measured short-run carry-over, not performance on a new unaided business task or months-later learning.[1]

The higher-minus-lower education gap on this separately standardized follow-up was 0.300 standard deviations in control and 0.200 among those previously assigned AI. The estimated difference in access effects was -0.100 (standard error 0.118; approximate 95% interval -0.331 to 0.131). The interval includes no change in the gap. The lower-education group had a positive follow-up effect, but a detected effect for one group and an undetected one for another do not establish that the effects differ. Separately standardized task and follow-up scores cannot yield a “share of learning retained.”[1][4]

An assisted answer need not survive the removal of the tool. Yet this immediate follow-up does not support a blanket claim that AI harmed unassisted performance here. Testing durable capability would require a fresh problem after time had passed, without AI.

A different experiment tests whether AI-assisted study leaves a benefit on later knowledge tests. At Middlebury College in spring 2025, 211 undergraduates attended a first lab session. Students were randomly assigned permission to use generative AI during a learning period on blockchain technology, carbon capture or CRISPR; those in the other condition were forbidden to use it. Both groups had reading material and could search other permitted resources. Both then faced knowledge tests on which external help was prohibited. Of the initial attendees, 204 returned for a second session after an average of 6.99 days.[2]

Assignment to AI permission raised the fraction correct by **6.7 percentage points immediately** (approximate 95% interval 0.43 to 12.97 points) and **5.1 points about a week later** (0.59 to 9.61 points). These are effects of *permission and access*, regardless of whether or how a student used the tool. The first test had five questions and the second ten, with different content and standard deviations; dividing the effects would not measure retained knowledge. The later positive estimate rebuts a universal prediction of no later benefit in this setting, while saying little about career-long skill.[2][4]

The intended-unaided scores require care. On a combined measure of first-session rule violations, AI permission raised violations by 12.6 percentage points (standard error 4.4 points), from a 4.9% control mean. Both tests barred external help, but the displayed integrity table gives no separate numerical compliance count for the second test. The paper’s estimate of how much first-session cheating explains depends on assumptions about violators and non-violators; random assignment does not identify that share. The high return rate and a check excluding students who reported studying between sessions help assess the later result, but do not verify that everyone complied. Nor does a fixed lab window show how students would allocate study time on their own.[2]

The comparison between experiments is about what each observes. Argentina tests *assigned AI access on a work task* across pre-existing education groups, followed by immediate related questions. Middlebury tests *AI permission while learning* among undergraduates, followed by a knowledge test about a week later. Middlebury has no less-educated comparison group; Argentina has no week-later test. Their different populations, tasks and horizons preclude a ranking of effect sizes or an education-based explanation for differences between the results.

Evidence	What the comparison identifies	What it cannot price
Argentine business-task experiment	Assigned AI access and scored output across pre-existing education groups; an immediate related follow-up	Degree completion, employer screening, sustained workplace output or pay
Middlebury learning experiment	Assigned AI permission during learning and intended-unaided topic tests immediately and about a week later	Education-group gaps, lasting career capability, hiring or pay
Historical US wage analysis	How education groups sort into occupations with different wage-growth paths	A causal AI-era change in the return to a qualification

From output to earnings

Assisted output can earn money without adding unaided skill. If AI stays available, a firm might pay a more productive worker more, hire more people, cut its price or keep the gain. Competitors and other workers can adopt the same tool. Learning offers another route: capability acquired with AI could improve later work, even when the tool is absent. A qualification may offer a third, through access to jobs and an employer’s judgment about whom to trust with work whose quality is hard to observe. Neither experiment estimates these earnings routes.

The gap in task scores is not the gap in earnings. In a snapshot, a group’s average rate of labour earnings depends on **which employment states its members occupy** and **what they earn within each state**. Weight each state’s current earnings rate by its share of the group; nonemployment has zero *current* labour earnings in this accounting.

That says nothing about a nonemployed person's earnings earlier in the year. A task experiment measures neither job access nor pay, and an observed difference between education groups is not the causal return to completing more schooling.[5]

David Deming's historical US analysis makes the job-allocation issue concrete. Following workers' employment and wage histories, he finds that college graduates enter professional, less routine occupations with more scope for wage growth over their careers. Early occupational sorting accounts for a substantial part of the widening observed college wage gap in his analysis.[3] That is a reason to look beyond a static test of task proficiency. It is not proof that a degree itself caused the sorting or that an employer's degree requirement is an effective measure of skill. Deming identifies several possible sources of the later wage growth, including learning on the job and better outside offers; his analysis does not separate them into a single definitive mechanism. It predates the AI changes at issue here.

Even a measured change in the college earnings gap would not be a net return to education. Fees, income forgone during study, the chance of completion and the timing of later earnings also matter. Argentina's education categories can both contain people who began tertiary study without finishing it. That makes a degree-premium calculation from this experiment especially misplaced; there is no defensible multiplier from its roughly 75% task-gap compression to wages.[1][5]

The direction of any earnings change remains open. If assisted output becomes easier to demonstrate, employers may hire capable people without a specified qualification. If qualifications still govern access to attractive jobs, or if educated workers also gain from AI, the task gap could narrow while the earnings gap stays put or widens. Even an unchanged earnings gap could conceal offsetting shifts in access and pay within jobs.

The observation that would matter

A more informative study would follow AI-induced changes in task performance **where actual hiring, job assignments and pay by qualification are also recorded**. It needs a credible comparison group and checks for changes in applicants, retention, demand and task mix. Identifying a link from AI-induced output to access or pay would support an earnings route; mere co-movement could not. A precise finding of no economically meaningful access or pay change over the study horizon would suggest that route was blocked or offset there. An imprecise null would settle little. A vacancy dropping its degree requirement would still reveal neither who was hired nor what they earned.

Longer follow-ups with new tasks and no AI would test whether assistance builds transferable capability. The Argentine email already shows that a tool can compress a measured task gap. To learn what happens to the value of a qualification, we need to see the employer's next decision: who gets the work, on what terms, and with what room to grow.

Sources

1. Guillermo Cruces et al., *Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment*, May 2026 manuscript, [arXiv:2608.04198v1](https://arxiv.org/abs/2608.04198v1), especially PDF pp. 3, 7–14, 18–19 and 31 (printed p. 30, Table 1).
2. Zara Contractor and Germán Reyes, *Experimental Evidence on the Learning Impact of Generative AI*, July 2026 manuscript, [arXiv:2607.08849v1](https://arxiv.org/abs/2607.08849v1), especially PDF pp. 6–12, 19–20, 38 (Table 4) and 41 (Table 7).
3. David J. Deming, *Why Do Wages Grow Faster for Educated Workers?*, May 2024 author manuscript, especially PDF pp. 1–5. The author's research page lists the work as forthcoming in the *Journal of Labor Economics*; the cited claims are from this manuscript.
4. Indiconomics calculation from the printed estimates and standard errors in Tables 1 and 4: approximate 95% interval = estimate $\pm 1.96 \times$ standard error. These are normal approximations, not participant-level re-estimation.
5. Indiconomics accounting framework, explained in [Methods and calculation notes](#); it is a conceptual identity, not an estimated model.

A27 methods and calculation notes

This is a comparative evidence synthesis. No new empirical estimate, meta-analysis, participant-level replication, college-return estimate or wage forecast is presented.

Different estimands

Cruces et al. randomize assigned AI access for one self-contained business task. Let H and L denote their pre-existing education categories and T assigned assistant access. The main interaction is $[E(\text{score} \mid H, T=1) - E(\text{score} \mid L, T=1)] - [E(\text{score} \mid H, T=0) - E(\text{score} \mid L, T=0)]$. This is a difference in access effects between education groups among analysed completers. It is not the causal effect of schooling. The score is scaled by the low-education control group's task-score standard deviation. The immediate related follow-up uses its *own* control-group standard deviation.

Contractor and Reyes randomize AI permission during a learning period for undergraduates. The reported fraction-correct contrasts are intention-to-treat effects of assignment on intended-unaided knowledge tests. Actual AI use, mode of use and study-time choices are endogenous; no treatment-on-the-treated or strategy estimate is used in the article. Their test had five questions immediately and ten about a week later. The percentages are percentage-point changes in the fraction correct, not a fraction of knowledge retained.

Deming analyses historical US worker panels and occupational sorting. His analysis motivates examining access to jobs and wage growth; it does not identify an AI-era response or isolate a signalling effect. The three studies do not estimate a common treatment, population or outcome, so we do not pool them or rank their effect sizes.

Mechanical reconstruction

Numbers are transcribed from Cruces Table 1 (PDF p. 31/printed p. 30) and Contractor/Reyes Table 4 (PDF p. 38); integrity numbers are from the latter's Table 7 (PDF p. 41). Approximate two-sided 95% intervals use estimate $\pm 1.96 \times$ reported standard error, without access to unrounded regressions or microdata.

Quantity	Calculation from printed table	Result / interpretation
Cruces main gap fraction compressed	$(0.548 - 0.139) / 0.548$	0.74635, described as about three quarters; no ratio interval available.
Cruces main interaction interval	$-0.408 \pm 1.96 \times 0.122$	-0.64712 to -0.16888 SD. The printed gap subtraction is -0.409 because of rounding.
Cruces treated-gap interval	$0.139 \pm 1.96 \times 0.087$	-0.03152 to 0.30952 SD; crosses zero.
Cruces follow-up interaction interval	$-0.100 \pm 1.96 \times 0.118$	-0.33128 to 0.13128 of the follow-up control SD.
Contractor/Reyes immediate ITT interval	$100 \times (0.067 \pm 1.96 \times 0.032)$	$+0.428$ to $+12.972$ percentage points.
Contractor/Reyes later ITT interval	$100 \times (0.051 \pm 1.96 \times 0.023)$	$+0.592$ to $+9.608$ percentage points.

We do not divide the follow-up gap by the task gap, divide the later student effect by the immediate effect, compare standard-deviation magnitudes across studies, or infer equality from a nonsignificant result. Required covariance information for a confidence interval on either ratio is absent from the printed table, and their denominators/outcomes differ. The confidence intervals are arithmetic approximations to printed standard errors, not a new uncertainty model.

Conceptual routes to earnings

The framework is unestimated. Current task output is $q = F(h, a, z)$, where h is human capability, a effective AI assistance and z task requirements. Cruces varies assigned access to a ; its education categories are neither randomly assigned nor exact observations of h .

Future capability can be written $h(\text{next}) = (1 - \delta)h + L(e, a, j)$, where δ is depreciation, e effort and j the learning setting or job. The sign of the AI effect on L is not imposed. Contractor/Reyes tests later topic knowledge under one fixed-time setting. Whether a person uses AI to augment or automate work is a choice, so use-pattern comparisons are not randomized learning mechanisms.

For a snapshot of observed group labour earnings, let D denote qualification status, A the AI environment, j a mutually exclusive employment state (including the relevant bundle for multiple-job workers), and Y the current rate of labour earnings. The accounting identity is $E[Y | D, A] = \sum_j P(j | D, A) \times E[Y | j, D, A]$. Include nonemployment as a category with zero current labour earnings; this is not a claim that someone currently without work earned nothing earlier in the year. It shows two margins: allocation across jobs and earnings within jobs. It does not identify the causal return to acquiring D . AI could alter output, its observability, employer screening, task prices, worker supply, demand and wage setting separately. Assisted output could affect earnings while unaided capability remains unchanged. Conversely, capability could rise without a pay or access change. Net educational return also requires fees, foregone earnings, completion probability, timing and discounting. No pass-through coefficient or net-return parameter is calibrated from these studies.