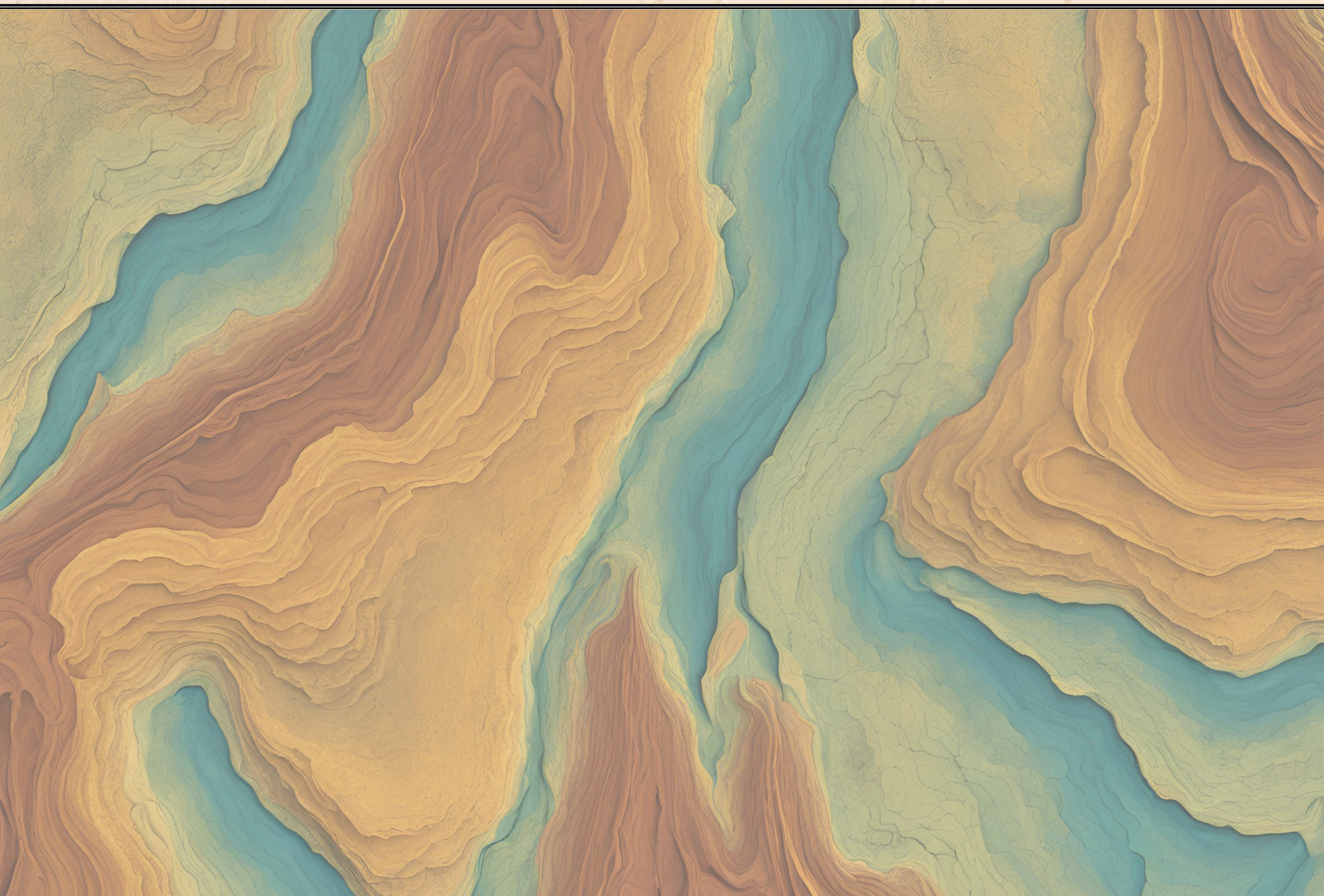
The Return to Top Schools

A review of the causal evidence

An Arb Research Report

November 2022



Executive Summary

- We look at the effect of going to an elite college instead of a merely good college, focussing entirely on studies with good causal identification. Most studies only look at the effect on income, and in the US or Chile.

- The formal evidence mostly points to small effects (0-10%) – but we're not confident in the methods they used. So it comes down to your prior about elite schooling. You can find a web app which allows you to apply your own judgment here.

- One measure of school quality (or selectiveness!) is the average SAT score of students who enroll. +200 points represents a large jump in school rank (hundreds of ranks).

- Applying our prior, using average intake SAT as a school quality proxy, we think:

Claim	Intervention	Operationally	Confidence
Income boost small (<= 10%)	Average to good school	1150 → 1350 SAT	75%
Income boost small (<= 10%)	Good to very good	1250 → 1450 SAT	65%
Income boost medium (<= 25%)	Good to top 10	1300 → 1500 SAT	60%

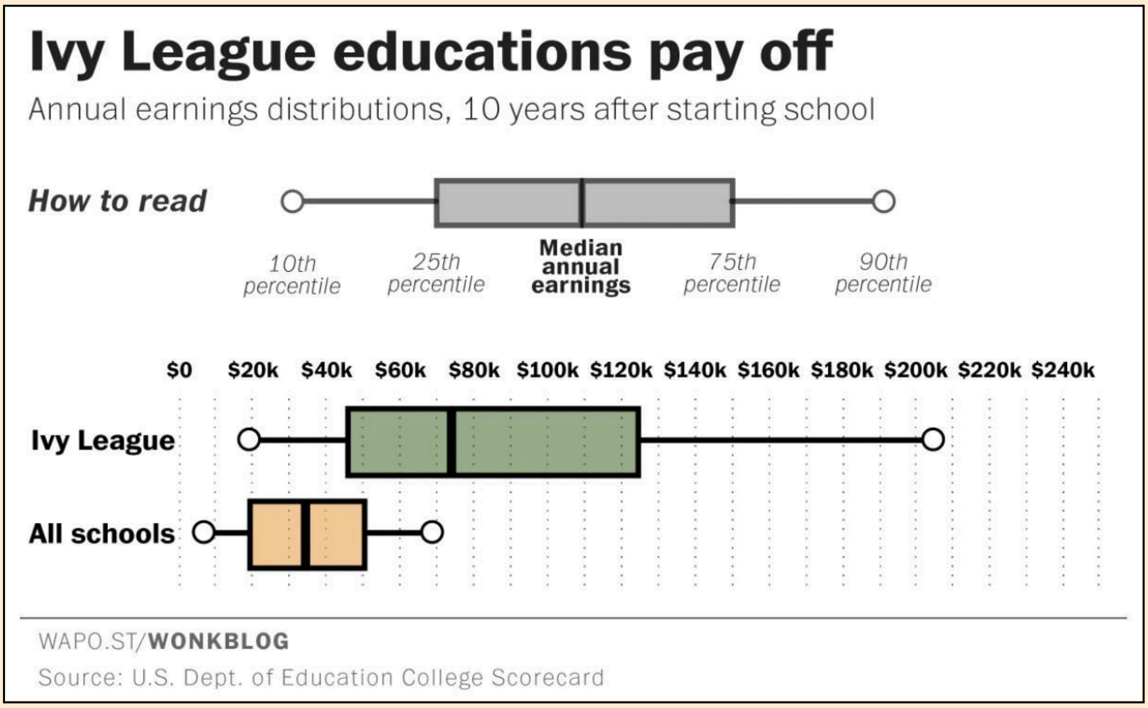
- Most of the evidence is about US students at US colleges; we are even less confident outside this domain. There is some evidence that effects are larger for ethnic minority students and very mixed for female students.

- This report represents 1.9 weeks of effort.

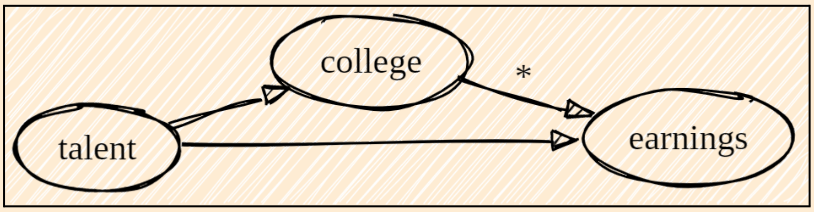
Why even ask this question?

There's a lot of emphasis on elite higher education — whether that's ambitious applicants deciding where to apply to increase their impact, or [funders](https://) thinking about what scholarships to offer. Economists are, surprisingly, less clear on the benefits.

On the surface, this is an odd thing to doubt. The 'effect' hits you between the eyes:



But this is a *paradigm* example of observational data misleading us: for instance, getting into an elite college correlates heavily with your talent, and talent correlates heavily with your earnings. (And this is just one confounder among many — consider also social class.)



✱ = the effect of going to an elite college, *given you're the kind of person who got in*

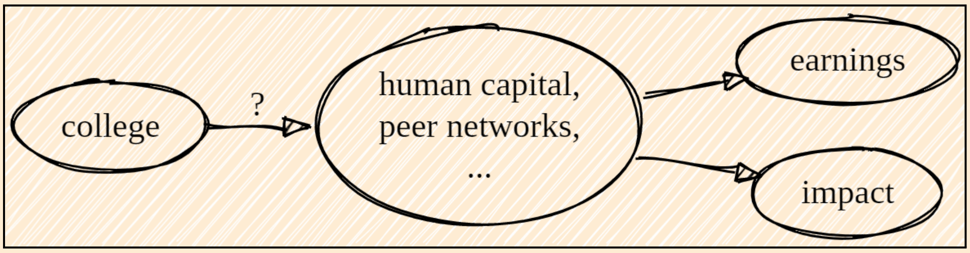
So, our research question

What's the *causal* effect of attending an elite college instead of a good college, particularly for the strongest students?[1]

We're interested in the effect of elite schools on subsequent social impact, but we can't see any studies on this. So we use any form of subsequent success as a proxy for added impact. (90% of studies use earnings as the outcome of interest.)

Why expect income to be a decent proxy for impact? Because some things about college cause both income and impact: e.g. knowledge, peer connections, first-job prestige and career tracks, or self-efficacy.

Income reflects your competence. Impact also reflects competence, and so if college affects competence then it should affect impact.



We demand causal identification, that the study be able to factor out unobserved confounders. This leads us to throw out *most* studies, because they just use ordinary regression plus covariates to "control" for obvious variation. This means they can't guarantee fully isolating the added value of attendance from the existing potential of admitted students ("ability bias" or "positive selection bias").[2] Briefly though: OLS studies find 10-30% increased earnings from elite college.

So: we want regression discontinuity designs (RDDs), twin studies, and (with some qualifications) instrumental variable and matched-applicant designs. (We briefly explain these methods in Appendix A.) But we should remember that all of these are essentially observational studies, underneath the clever tricks which may somewhat reduce bias.

[1] Note that our question is very different from the *average* private return to education in general (i.e. pre-college schooling), which seems high, particularly for disadvantaged groups. It's also different from the social return, memorably and eccentrically analyzed in Caplan (2018). Note also the surprising evidence for the causal effect of high-school quantity on intelligence.
[2] Some economists disagree, and say that in their samples, with their covariates, the selection bias looks manageable. Long (2010), Borgen (2014).

Initial inclusion criteria

- Good causal identification
- Effect of elite college vs other college
- Any form of private return

Initial results

The resulting rapid review is in Appendix B. Despite our requirements being really specific, we still found dozens of studies. As usual, the results are a mess:

- +33% income for going to the elite college in Milan.
- No effect in the US.[3]
- +15% MBA earnings in the US per 100 point increase in school average GMAT.
- No effect on earnings in Texas per +100 points in average school SAT. How to manage this?

Refining the search[4]

1. Omit studies from undifferentiated college systems

Countries differ by how much their colleges differ. For instance, within the economics of education, Australia has famously flat college quality.

We could do something principled by looking at the spread of college quality within countries. In practice we'll just focus on the US, UK, Chile, and China, as instances of very differentiated systems.

But even the US, with its huge spread in quality and perceived quality, is less clean-cut than places like Chile, where the consensus is that the top 2 colleges are much better than the others. This makes it harder to apply Chilean results to the US.

2. Downweight studies about less elite (top 5%) colleges

Most studies understandably focus on the average effect of going to college at all, since this is the practically important domain for most people. But this isn't what we're trying to study.

There is a separate literature on "college quality", which tries to look at the effect of better or anyway more selective colleges. But even this sub-literature can be too permissive for our

[3] Or if you look at the original preprint, +10.8% in one of the dozens of analyses.
[4] We came up with these three extra criteria after seeing the papers, and this introduces a researcher degree of freedom, slightly undermining our results.

question (e.g. they look at the top quartile of quality, where we are most interested in the top 5% or 1%).

At present, the average SAT score is 1060 and the average intake SAT of the Ivy League is 1506. So we can make use of studies which use SAT as a proxy for college quality by taking +450 units, which simulates the effect of switching to a top college. (In the 70s, the focus of some big studies, this difference was more like 300 points.)

Similarly (under the assumption that EAs applying for elite scholarships are above average) it's ideal to have outcome quantiles instead of a mean or, worse, median outcome.

3. Downweight studies which use early-career outcomes

Early career income (i.e. before age 30) is a very noisy predictor of lifetime income. This is partly due to first jobs being atypical (think Teach for America, law clerks, or the Peace Corps, which have much lower pay).

But more: many economists think the premium to elite colleges actually disappears after the first decade or so of work (so-called "employer learning"). So we should measure the effects well into the student's career - ideally lifetime earnings, but the most extensive studies to date cap out at 30 years after admission.

4. Downweight studies with only one measure of college quality

The magazine US News publishes one of the most influential rankings of US colleges. An investigation of his own institution's metrics by a Columbia mathematician found clear evidence of "goodharting" (distorted institutions, measurement, and reporting to fool a metric) at the college. In fact this is so intense that Columbia's current US News ranking of #2 could actually be #12 or #26 in the absence of goodharting. We expect the college to be unusual only in the degree of corruption, if that, and other rankings to suffer distortions if their covariates are similar to those of US News.

If every college lied the same amount, then we could relax and take the ranking at face value. But this isn't a safe assumption, so we recommend looking for studies with several measures of college quality, with different methodology. This extreme measurement noise probably explains some of the incoherence in studies of the US.

Takeaways

- The formal evidence suggests that the average effect of elite undergraduate education on US **earnings** is **small**, after a 10+ year followup (**0% to 11% increase**).
    - The definition of "elite" used is inconsistent, but the distribution roughly centers on moving from a top quartile school to a top 2% school (top of the country studied).
    - Some evidence that this washes out over the first 15 years of the career.
    - Some evidence that this premium is less than half of its 1980s level.
    - Generally larger effects in low/middle-income countries.

We can use the naive effect (i.e. the average income change, without controlling for selection bias) to get an upper bound on the causal effect.

10 years on, the median admission to the 'top' college (per US News) can expect:

- +146% earnings above the average college.
- +65% above college #100
- +51% above college #30
- +12% above college #10

(If we go by our review of causal estimates, more than 90% of this is bias.[5])

- Decent evidence that the average effect of elite education on **earnings** is 5% to 15% for minorities and first-generation students.

- Some evidence for a 2% to 20% effect of attending elite college on the probability of **completing college**.

- Weak evidence that the effect on **top 1%** students is **large** (30%+ effect on earnings, probability of extremal earnings, leadership positions).

[5] This claim, of 90% bias, implies that students, application advisors, and/or college admissions are *extremely* good at predicting talent or at least earning potential, since we end up with an allocation with a good approximation of the rank order.

How can the causal effect on earnings be small?

This is quite surprising! In the *Washington Post* graphic at the start of this post, we see a massive naive effect of elite college: doubling income. This accords with the common-sense view about these schools.

Further, 1) students and their parents have strong preferences about going to better schools, and devote years of resources to getting into them; 2) people seem to pay attention to education when evaluating others; 3) there are several clear mechanisms which should lead them to improve career outcomes. (Consider human capital accumulation, strong peer networks, assortative mating, and job market signalling.
Befriending smart people really should be helpful, because of tips and referrals!) But we see

that controlling for selection bias shrinks the effect enormously.

What could explain the minor effect of elite college on earnings? Some combination of:

1. *Very likely*: ability bias (the effect by which people with higher existing earnings potential sort themselves into better colleges, inflating the naive estimates of the college effect). For average colleges, estimates range from 10% to 40% of the naive income effect. Our review of elite college studies implies it might be even higher for the elite case.

2. *Very likely*: large non-monetary payoffs. We might infer this from cases of parents paying bribes in excess of the expected financial return to their children.
    a. *Access to power*. An unusual number of elite grads go into government and academia, with much lower pay but high social impact.[6] 35% of Caltech graduates immediately go to grad school. For someone pursuing an academic career, the prestige of where you get your *doctorate* seems crucial. This is probably ability bias again, but for fields where judging candidate quality is hard this might still pay off.
    b. *Partner quality*. One study finds a small positive effect. One study of education quantity in Denmark found that 50% of the return to general education went through marriage effects.
    c. College as *consumption good*. It's fun for many students, and by revealed preference it's valuable to be associated with elite colleges.
        i. People tend to enjoy (or appreciate) their college years despite not learning as many practical skills as they would at a good bootcamp. People often buy experiences, often seduced by peer pressure.
        ii. Similarly, more selective colleges are prestigious and people are willing to buy more prestigious things (e.g. Rolex) and associate with prestigious institutions (e.g., donations to the vast Harvard endowment) even if it brings them no material benefits.

[6] e.g. Since Churchill, 11 out of 14 UK PMs studied at Oxford. See also US presidents. Then consider the tight connection between top schools and power in the UK and France.

   d. *Status*. College prestige often serves as a proxy for productivity and impressiveness.
      i. Employers presumably learn about worker productivity through job performance. But many social interactions are brief. Having an Ivy on your resume might not add much value for (yet another) career transition, but it may still impress some as an offhand remark at a party or a dating profile.
      ii. Further, for some activities (like being a public figure), people are not motivated enough to evaluate carefully (or are not capable of such careful evaluation), so they would fall back to prestigious proxies.

3. *Likely*: "Major > school". One line of thought emphasises *what* you study instead of where you go. (This could be relevant because elite students disproportionately go for social science degrees.) Good college quality studies adjust for this, e.g. Anelli (2016), Andrews (2012), Mountjoy & Hickman (2021), with inconsistent results.
   a. Mountjoy & Hickman (2021), the most credible study, finds that "students who attend the more selective colleges in their admission portfolios [...] are significantly less likely to successfully major in STEM". This is consistent with the intuition that doing, say, English at a more demanding school might be easier than doing math at a less demanding one. At the same time, "*more selective colleges tend to have lower value-added on STEM degree completion.*"

4. *Plausible*: 'employer learning' about worker productivity. There's a surprising consensus among economists that the elite premium washes out (i.e. that there is a diminishing instead of compounding wage premium) throughout the early career. This is surprising because it contradicts intuitions about "tracking" and path dependency from first jobs. Bordon & Braga (2017) find that the elite college premium disappears after 6 years (in Chile). Démurger et al (2019) find that the premium decreases by an order of magnitude after 15 years (in China). Mountjoy and Hickman (2021) see this in Texan premia after 10 years. In contrast, Borgen (2014) finds an *increasing* return to college quality in Norway. Overall we're not sure about this and haven't vetted this claim very hard.

   One possible story: "if you have a degree from a top-school, you might go to a top job directly, while undermatched kids first take a quick stop at another college. After 10 years, they are both L5 engineers at Google."

5. *Less likely*: "Peer networks are less important than they seem" / "peer networks are permeable after college" / "social class (i.e. family network) is more important for income". These are less likely because, despite strong selection for low socioeconomic status, only 15% of Harvard students are first-generation college students and we see larger effects of college for them (see Appendix B).

How we arrived at our conclusions

We begin with the prior that there should be some causal effect on earnings, possibly large, given the intense investment by many different parties with a strong interest in getting it right (students, parents, employers).

We find a low college quality premium in an array of RDD studies, in one great matching study, and in several less good matching studies. We discard OLS studies, justified by the alarming result in Mountjoy & Hickman (2021) that even including very rich covariates about students doesn't eliminate selection bias. We weight the studies by data and method quality, and by relevance to our question (US, most selective colleges, 10+ year followup). Gavin puts 67% of the weight of evidence on studies finding small or null income effects (i.e. less than an 11% increase). This is then set against a fairly strong prior in favour of elite education.

There's good evidence of publication bias in nearby literatures. So, there may be missing null effects. On the other hand, publication bias often filters for *surprising & socially desirable* things rather than just $p < 0.05$, so a null for elite education is far more publishable than other nulls.

Several studies reported the premium decaying over the career (under the name "employer learning"). Only one study found the reverse, in Norway. We are agnostic about the period of this decay (though we focus on 10 year follow-ups, and the income effect is apparently small by then).

Effects also tend to be smaller in recent cohorts and in richer countries.

Some medium-quality studies find some effect on the probability of completing college, and we don't find any results contradicting this.

Despite their problems, we decided to highlight results about top percentile outcomes, since other studies generally look at average effects, and there might be such an effect, at the elite scholarship kid level.

We sanity-checked all of this using studies of elite high schools. This also served a negative result about ultra-elite schools (average GPA at the 99th percentile or higher). This one is a regression discontinuity, and thus looks at the worst top-performers - so the study is informative about the right tail, but not as much as one might assume. Studies of elite high schools generally agree that, after controlling for selection bias, the causal effect of these schools is small in the US.

Surveying alternative explanations for the large investment in college applications, we find them adequate.

Overall,

- We're 75% confident that for a marginal student the income effect of attending a 1350 SAT vs. 1150 SAT school is <11% in the US.
- We're 65% confident that for a marginal student the income effect of attending a 1450 SAT vs 1250 SAT. school is <11% in the US.
- We're 40% confident that for a marginal student the income effect of attending a 1500 SAT vs 1300 SAT. school is <11% in the US.
- Note that every subsequent estimate has less and less credence: we are pretty sure about average schools and much less sure about more elite ones.

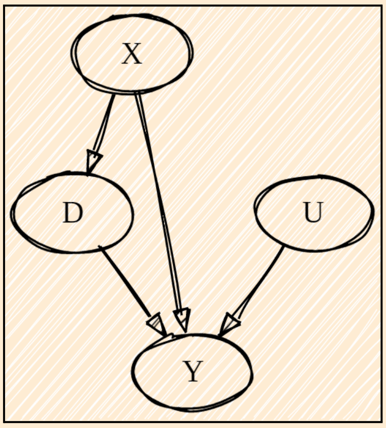
See also

- 80k on the average case.
- The disproportionate amount of seed funding given to elite graduates (not causal evidence).
- Good frontend to the US college data

Thanks to Phil Harrison, Sam Enright, Alvaro de Menard, and Stag Lynn for comments.

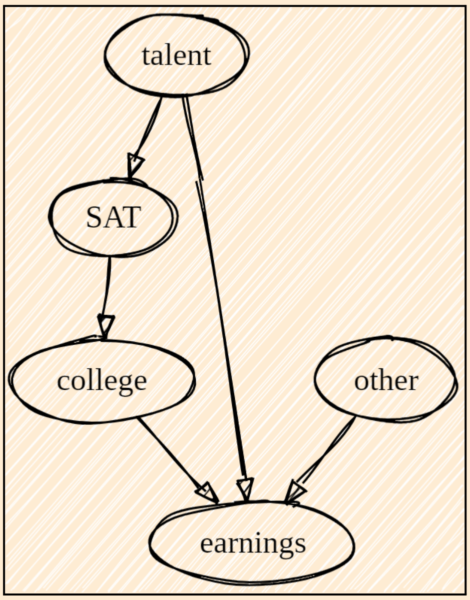
Appendix A: Just enough econometrics

In general we have the causal graph:



We know some things about our participants, X. We want to know the effect of a treatment D on some outcome Y. But it's confounded by X. And we're not randomizing the assignment of individuals to D, so [unobserved factors] U might be crucial.

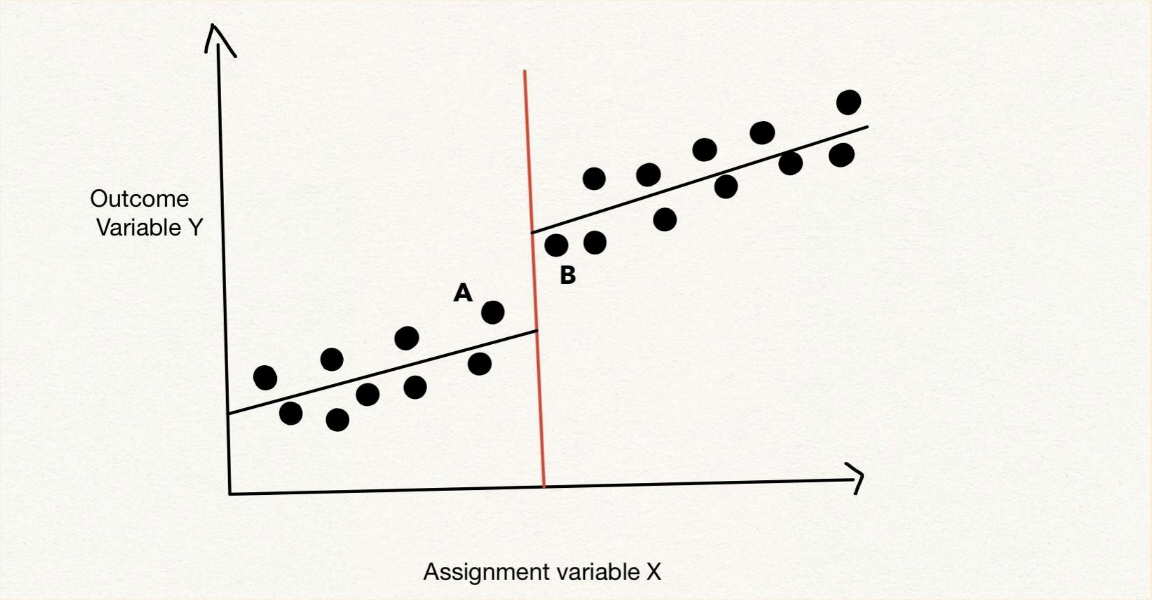
In our case:



We don't have RCTs and probably never will. What causal inference can we do?

Regression discontinuity design (RDD)

RDDs are good in principle but are often misused (see “Problems”).



A = outcome for participants just below an exogenous cutoff B = outcome for participants just above an exogenous cutoff Estimate = B - A

We want to reduce the confounding from omitted variables U, but we can’t do this in the proper way, by randomizing assignment to D. However, sometimes discrete processes (e.g. government rules or disasters) randomize people.

Take some continuous measure of ability, e.g. academic potential as proxied by SAT score (X). Then find situations where a cutoff only assigns students to the treatment D if they have X above a threshold. Then look at students *just* either side of the cutoff. They should be basically the same, and thus can be treated as if randomized to D.

In our case the exogenous assignment comes from national entrance exam cutoffs to be eligible for elite colleges, or eligibility for financial aid, etc.

Problems

- The design means that the *treatment assignment* only depends on X. But to obtain unbiased estimates, you will still need to adjust for all pre-treatment differences between the groups, and this is hard to know and often hard to collect data for. Why?:
    - It is likely that other informative features will co-vary with X.
    - For test scores, one such feature might be parental income. If you naively do two linear regressions as in the diagram above, without adjusting for parental income, then you might compare groups that differ importantly.
    - For example, if parental income changes nonlinearly with test score, the group above the cutoff might have spent much more on tutors, and we wouldn't be comparing apples to apples.

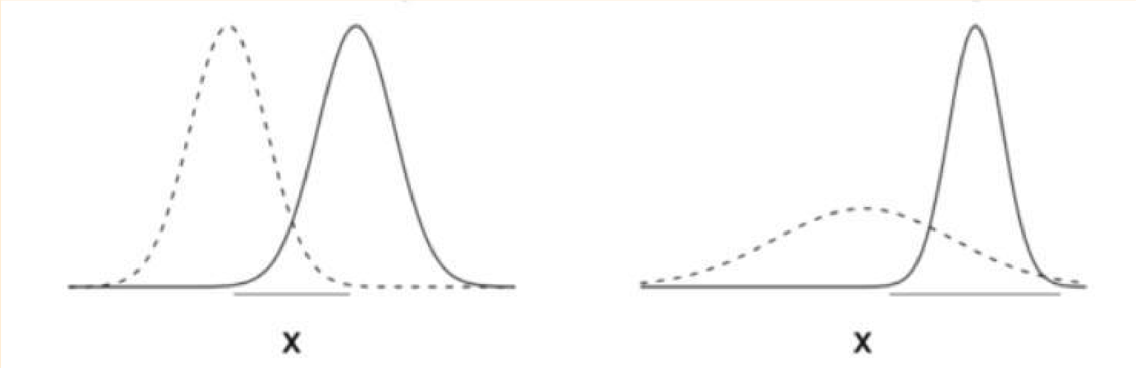
- It is also critical to check the continuity assumption, that in fact people either side of the cutoff are very similar.

- The slope tests used often lead to nonsense getting published. It is very important to eyeball the resulting graphs. Here's a useful checklist.

- *By construction*, RDDs look at the worst top-performers. This is usually fine, but not for our research question. We can get some mileage out of these studies if they include a sensitivity analysis varying the "bandwidth" (distance from threshold).
    - At the same time we could be interested in marginal applicants, since they could use help more. But it's plausible that marginal people moved by a scholarship are different from the marginal people affected by SAT score.

Matched-applicants model

This design is less strong than RDD, since it doesn't involve any randomisation.



From [here](). Consider a case where the distributions of the control (dashed) and treatment (solid) partially overlap. The underline denotes the Xs that get matched.

Students who ended up in very selective schools and average schools are pretty different, so we can't compare these populations directly to understand the average treatment effect of attending more selective schools.

Matching is a technique to make the "distributions" of students in control and treatment groups more alike. More specifically, matching selects a subset of the data with "complete overlap" and then discards everything else: the subset of the treatment group for which there were no similar points in control, and vice versa. Matching may further require some weighting, giving some observations more weight, to make distributions more alike. (This should be distinguished from the more common propensity score matching method.)

In our domain we often match *applicants to applicants by admission profile*. Say students A, B, and C applied to the same three schools and were accepted and rejected by the same/similar schools. We:

1. Match them as sufficiently similar, creating an indicator variable for them alone.[7]
2. We stratify the analysis on these indicators.
3. Reweight. Say that A and B went to a more competitive school and C to a less competitive one. Then in the aggregate we will give C twice as much weight as A and B.

Comparison with regression discontinuity:

- In a way, the two methods are very similar: let's take sufficiently similar distributions of students who went into different colleges and calculate the average effect over these.
- RDD requires a coincidental cutoff, which leads to random assignments around this value of X: a close-to-cutoff student could have gotten lucky and gotten in, and the unlucky didn't.
- Matching doesn't employ randomness, but it makes the distributions sufficiently similar by only taking points that look similar.

Problems:

- No randomisation. Take two colleges, with one better A > B on some measure (e.g. average intake SAT). A student choosing B is a signal about B (or their fit with B). They may be more interested in a specific course, or sports team, or professor at B; other factors like financial aid or proximity to home might be decisive; or the SAT proxy may have erred in this case and B is actually more prestigious.

  As such, the method is not even guaranteed to *reduce* bias relative to unmatched OLS.

- Mountjoy and Hickman (2021) note that these designs "*sacrifice some external validity on the altar of internal validity by limiting the identifying variation to enrollment choices within admission portfolios*". That is, we can't identify other effects across categories of college (e.g. 4 year vs 2 year) or across types of student.

- Some papers match on "similar" school sets rather than identical sets. If this is done naively, e.g. by school SAT, this can produce matches on implausible sets of schools.

Self-revelation model

Confusingly, “matched-applicants” actually matches based on colleges (the colleges which admitted a group of students). A related technique, “self-revelation”, instead matches students based on costly signals from their applications. This design is not much better than OLS, despite it being central to the most prestigious papers.

Dale & Krueger use the average school SAT the student applied to, and the total number of colleges they applied to, as a proxy for their ambition or "unobserved earning potential". “Average school SAT of the set applied to” seems quite strong: the student should have some information about themselves and it's very costly for them to bluff. (That is, if a student applies to colleges well outside their reach, they run the risk of not being accepted to any of them. This large cost should make applications a good proxy for talent.)

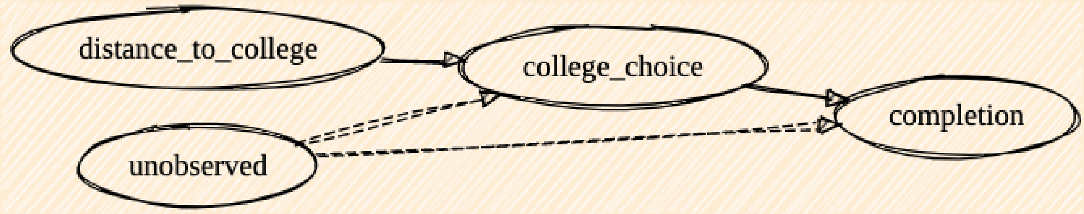
These are simply extra covariates which adjust student returns by another proxy of talent besides test scores.

Problems

- No randomisation. As above, but more severe since we’re using a less stringent method and using it without explicit matching.

Instrumental variables

In principle this design is somewhere between matching and RDD in causal identification, but in practice many published instruments are weak.



Instrumental variables (IV) estimation uses a variable (e.g. distance from home to college) that induces changes in the explanatory variable (e.g. college choice) but has no independent effect on the dependent variable (e.g. college completion). This allows researchers to extract causal relationships by comparing subgroups: treatment members where IVs are present, and control members where IVs are not present.

[Cohodes & Goodman (2014)](#) use proximity to a flagship college as an IV: proximity might swing an applicant to attend a college, but supposedly doesn't affect completion probability, independent of that.

Problems

- Causal identification is here (really) an assumption, hence a theoretical justification is crucial.
- An instrument (distance to college) might be correlated with omitted variables (unobserved), which defies the premise.
- If an instrument is only weakly correlated with the endogenous regressor (college choice), we know that estimate will be biased in small sample sizes.

Appendix B: The Rapid Review

Causal evidence for elite college premium

Keywords: "elite college attendance", "college quality", "quality premium", "selectivity premium", "match effects", "university prestige", "undermatching"

Initial inclusion criteria

- Good causal identification
- Effect of elite college vs other college
- Any form of private return

Post-hoc inclusion criteria

- Study of a differentiated college system
- Top 5% colleges
- >10 year followup

Desiderata

- Top-percentile students ideal
- US evidence best
- International student evidence ideal
- *Graduation* (rather than any attendance) would be good
- Recent work better (external validity + method improvements)
- Or long-term (30 year) followup

Limitations

- Undergrad focus.
- No focus on international students
- Mostly look at attendance instead of graduation, and there's a well-known discontinuity in the effect for completing.
- Massive increase in tuition costs (and debt) since many of these studies
- We don't separate relative (being at the top school in your country) and absolute effects (being at a better school). You could call the two effects 'selectivity' and 'quality', but the literature mostly doesn't distinguish the two.
- One line of thought suggests that we should use "college quality" to refer to the value-added of the college, rather than the current meaning, prestige / input intensiveness / tuition cost proxy.
- We don't investigate whether effects differ for different measures of college quality (average intake SAT, costly inputs per student, reputation rankings, etc). Some suggestion that school SAT is a poor predictor of earnings.
- Evidence of publication bias (against null studies, as usual). So even our small effect

might be an overestimate.
- We don't investigate how good income is as a proxy for impact.
- Most results don't give us a *return* (benefit net of costs and opportunity costs), just a gross benefit.

Bold authors = passes all inclusion criteria
Most earnings coefficients are reported as additive increase in natural log of earnings per unit college quality (for instance, per 100 points of average SAT in the intake class). I've rescaled them to *% increase in earnings from average college to elite college*.

We aggregated the results in Weighting the studies
Total weight on a small (<11%) effect: 67%.

Appendix C: Deep dives

We now look closely at a few studies, selected by relevance and quality. TL;DR:

- Dale & Krueger (2002)
    - Notable for: innovative adjustments for selection bias, being the locus classicus.
    - Result: headline is "null effect on earnings," but results are actually inconsistent across the different college quality proxies used.

- Dale & Kruger (2011/14)
    - Notable for: a longer followup on the classical study, administrative data.
    - Result: same as above.

- Zimmerman (2019)
    - Notable for: RDD looking at ultra-elite outcomes with a 30 year followup.
    - Result: attending the top 2 Chilean universities (i.e. top 3% in the country) leads to +44% probability of being in a leadership role (director or C-level) and +51% probability of being in the top 0.1% earnings *for one subgroup (while no effect in other 3; and no pre-registration)*.

- Andrews, Li, & Lovenheim (2012)
    - Notable for: quantile analysis, showing a large spread for the effect on earnings, with large effects at the top few percentiles of earnings.
    - Result: +32% earnings for men at 97th percentile of earnings, after 5-13 years.

- Mountjoy & Hickman (2021)
    - Notable for: great identification analysis (via matching by the set of institutions an applicant was admitted to) allows them to address selection bias and evaluate the value-added of each college across different students
    - Result: basically no effect on +100 avg. SAT intake on earnings and on not going to UT Austin, in the world top 100.
    - Takeaway: benefits vary a lot across colleges — and while selectivity is a nearly perfect predictor of raw income, it is a poor predictor of the *value-added,* after accounting for the colleges students are admitted to.

- Dobbie & Fryer (2014)
    - Notable for: RDD on NY high schools, average SAT at 99% to 99.9% percentile.
    - Result: little impact on the cutoff students' college enrollment, college graduation, or college quality.

Dale & Krueger (2002)

- *Population*: US college attendees, 1976 cohort. Includes elite schools like Yale, Swarthmore, Columbia, Rice, Princeton, Penn, Stanford.
- *Intervention*: Attending an elite college / not attending after admitted. Matched applicants, but using "equivalent" admitting schools (instead of the same admitting schools) to overcome a lack of data.
    - *College quality measure*: 3 different proxies!
- *Control*: Applicants who got into a college but didn't attend.
- *Outcome*: Self-reported income.
- *Time*: 19 year followup
- *Sample*: n = 6,335. Convenience sample, response rate ~80%.

- Novel methods: Matched-applicant and self-revelation. More or less clever ways to restrict the data to a *potentially* less confounded subset.

Result

The headline the authors prefer is a **"null effect" on earnings** when using school SAT as the college quality measure, and using self-revelation to control for selection bias.
However, there are actually 24 analyses in the paper: some results are ~0 and some are large. The strange default unit used — "additive increase in log_e(earnings) per 100 points of average intake SAT" — has also misled most commentators.

One good thing about the paper is that it uses quality measures *besides* average intake SAT. Pulling out and scaling the relevant numbers shows a range between +4% and
+11% earnings.

SAT as quality proxy

- Table VI: including self-revelation variables, we see +1.3% *log* earnings per +100 points of average intake SAT.
- In 1976, the average SAT was 1006, and an elite school looks like 1300+.
- So for elite schools, exp(1.3% * 3) = **+4% earnings**. (Large standard error, hence the frequentist headline "no effect".)

Barron's index as quality proxy ("most competitive")

The original preprint finds an effect (Table 7) for attending the most selective category in the *Barron's* ranking (roughly the top 50 US schools). **+11.4% earnings** for men and
+6.9% for women, using matched applicants. This analysis ("most selective" alone) is missing from the published version. This seems the most relevant for our question.

Tuition as quality proxy

"an internal real rate of return of 13% [in 1976 or]… **8% [in 2002]** if tuition costs are doubled". Note that this is a much better outcome than +8% earnings! But not totally relevant for OPP scholars.

Limits

- No pre-registration
- Convenience sample
- *Matching is still observational*. Zero randomisation. And the decision to enter a less prestigious college is clearly nonrandom.
    - Don't have info on student's financial aid offers, a powerful explanation for differing college choice.
    - Doesn't match on gender
    - Doesn't seem to adjust by major

- Self-reported income is a "soft" measurement (open to overreporting), but has some advantages over administrative data (which must omit the large tax avoidance).
- Overheated critiques here and here. A few of the D&K analyses yield large effects, but they may have been dropped because of banal coding errors rather than storyline management. If average intake SAT is shown to be a poor proxy for college quality in this period, then we could go with the preprint Barron result.
- Only n = 2,330 exact-school matches, and Avery & Hoxby (2003) critique the similarity of schools in the matching they used instead. "*Dale and Krueger were forced to merge colleges into crude "group colleges" to form the cells. However, the crude cells made it implausible that all students within a cell were equal in aptitude, and this implausibility eliminated the usefulness of their procedure.*"

Dale & Krueger (2011/14)

The same authors repeated their analysis with a longer followup on the same data, using much the same matching techniques.

- *Population*: US college attendees, 1976 cohort and 1989. Includes elite schools like Yale, Swarthmore, Columbia, Rice, Princeton, Penn, Stanford.
- *Intervention*: Attending an elite college / not attending after admitted.
    - *College quality measure*: 3 different proxies!
- *Control*: Applicants who got into a college but didn't attend.
- *Outcome*: Administrative data on income (rather than self-report).
- *Time*: up to 30 year followup
- *Sample*: n = 14,238. Convenience sample, response rate ~80%

- Ge, Isaac & Miller (2019) is essentially a reproduction with the same data.

Result

Again, the authors prefer the null result, out of their many analyses. "*when we partially adjust for unobserved student ability by controlling for the average SAT score of the colleges that students applied to, our estimates of the effects of college characteristics fall substantially and are generally indistinguishable from zero*".

Except the **+5%** to **+15%** Black/Hispanic **earnings** across all college quality measures (Table 8).
And the **+8%** first-generation college student **earnings** (Table 9).

As before, we can look closer at the tables for our purposes:

Average intake SAT as quality proxy
Table 5: Using the last income report: 2003 earnings through 2007. With self-revelation adjustment: -0.008 * 3 = **minus 2.3%** earnings

Tuition as quality proxy
Unclear. Table 6: 0.026 * some log increase in tuition.

Barron's index as quality proxy ("most competitive")
Table A5: **+21%** earnings for the 1976 cohort, or **+11%** for the 1989 cohort, using self-revelation as adjustment for selection bias. Large standard errors.

Limits

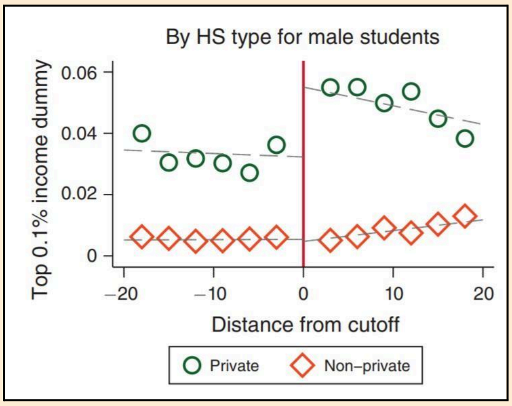
- Not pre-registered. (Limited degrees of freedom though, since it follows prior analyses.)
- Matching is still observational. Zero randomisation. And the decision to enter a less prestigious college is clearly nonrandom.
  - Don't have info on student's financial aid offers, a powerful explanation for differing college choice
- One problem with using administrative income data is that it doesn't include the 20% of income rich people manage to avoid tax on. It also seems to ignore capital gains income.
- Just income (see above).
- Critique, not very relevant
- Scott Alexander takes the headline result on face value and doesn't consider the non-monetary payoffs, but he comes to otherwise sensible conclusions.

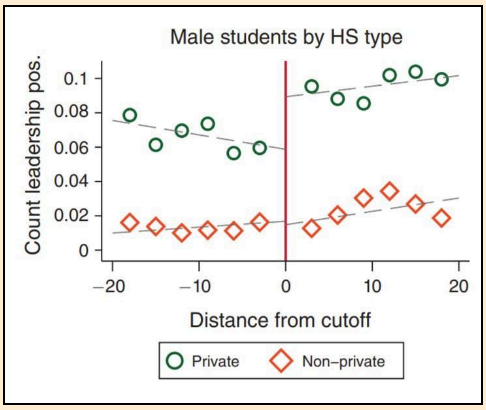
Zimmerman (2019)

This paper is unique in looking at ultra-elite outcomes and has a 30 year followup period. It uses a regression discontinuity about the Chilean national exam threshold.

- *Population*: Chilean business majors, 1980-2001
- *Intervention*: "Randomised" to the top 2 schools (above cutoff).
- *College quality measure*: top exam score bracket
- *Control*: "Randomised" to other Chilean colleges (beneath cutoff)
- *Outcome*: eventual leadership roles or extreme earnings. Admin data
- *Time*: 30 year followup
- *Sample*: n = 2,522 and 36,211. Complete administrative coverage of population!

Attending the top 2 Chilean universities (i.e. top 3% in the country) leads to
- +44% probability of being in a **leadership (director or C-level)** role
- +51% probability of being in the **top 0.1% earnings**





Limits

- Business majors only
- RDD: by definition, not looking at the top students
- Chile is one of the most differentiated systems in the world - large gap between the top 2 colleges and the others. This makes it hard to infer things about less clear systems like the US, where the ranking at the very top is contentious.
- There is presumably some effect from relative quality ("best in Chile") and some from absolute quality (world #100 college vs world #1 college). These elite Chilean universities are top 200 in the world but not top 100. These specific outcomes probably reflect the relative national rank more. But reasonably, some effect from moving to #200 to #1 as well as #400 to #200.
- "income measure omits business profits that are reinvested in firms, which may lead to underestimates of top income shares"
- Subgroup analysis: **0** effect for women, **0** effect for state high-school students
- Given no pre-registration, and the result holding in only one double subgroup (privately-schooled men) we should be dubious.

Andrews, Li, & Lovenheim (2012)

This is just a fancy OLS study ("unconditional quantile treatment effects"). Normally we exclude OLS studies, but we make an exception for this one because it points at the spread of effects *within* elite students: "*the quantile treatment effects are outside the mean confidence interval for over 70% of the distribution*".

- *Population*: male Texan public university graduates. 1996-2002 cohorts.
- *Intervention*: **graduating** from UT-Austin
    - *College quality measure*: eyeballed
- *Control*: other Texas schools
- *Outcome*: administrative earnings data (rather than self-report)
- *Time*: 5 to 13 year followup
- *Sample*: n = 9,837 UT-Austin grads, n=94,071 all grads. Administrative data with complete coverage, but filtered by staying in-state for school and subsequent work.

- Going to UT-Austin makes you top 3% of students; then within that top 3%, this paper tries to say something about the top 3% earners.
- The subgroup, male students, was chosen due to women being missing from the earnings data for endogenous reasons (fertility?). Probably not too problematic a forking path.

Briefly, they use attending UT-Austin as their 'treatment' and try to construct a synthetic control group by using the observed characteristics of these students. Call treated students A. They produce 99 estimates, one per percentile, attempting to produce a "counterfactual" distribution (the outcome for A had they not attended UT-Austin).

You can then get a distribution of effects by doing (UT-Austin effect minus other-college effect) for each quantile of the student distributions. They argue that the key assumption holds (that the covariates are enough to control for selection bias), but see below for reasons to worry.

They find that the income effects of UT-Austin (and other tiers) vary hugely.

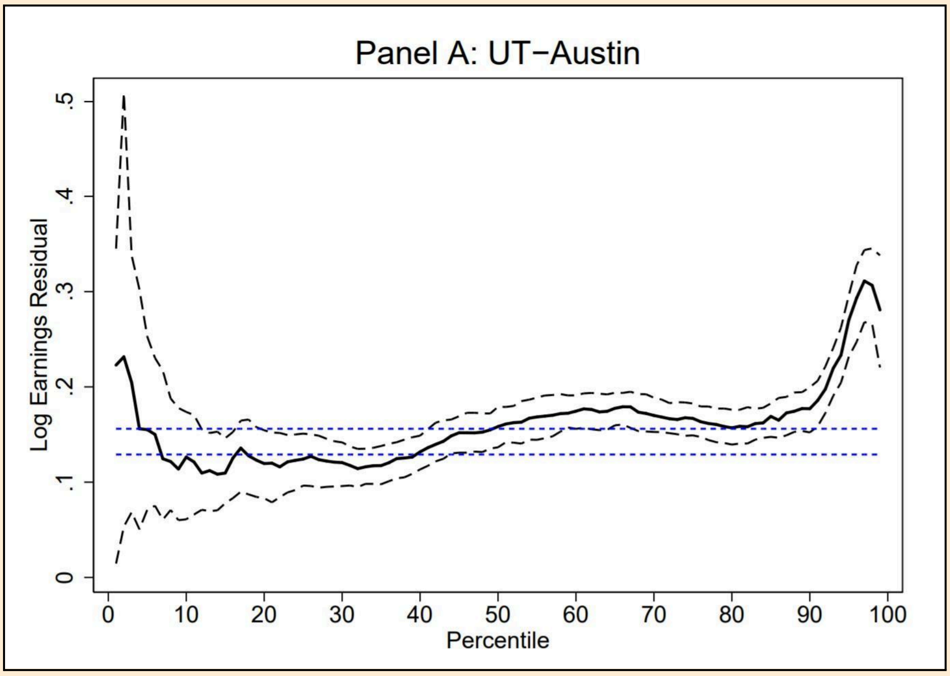


Fig 6A. Blue lines are baseline and dashed black lines are the confidence interval.

Result

+32% **earnings** for men at 97th percentile of earnings, after 5-13 years. Or, extracting the relative effect we can trust a bit more: the effect of elite college is **3 times higher for the highest-earning** graduates. Incomes follow a power law - no surprises there. But this is an order of magnitude difference in the *change* in income, within the treatment, hidden behind the average. This could be important.

Previously, we noted large selection bias in OLS studies, such that estimates are much higher than causal studies. But, for this paper, we're not interested in the absolute estimate but instead the variance in it. 9th percentile incomes at the elite college were +2.7% above non-elite college; but 97th percentile incomes at the elite college were +31.7%.

Limits

- Not causal evidence. But we're not taking the absolute level literally: instead we're treating it as evidence that effects could vary by an order of magnitude within the top of the top. It also makes our critique of RDDs (as measuring only the worst of the best students) concrete: if the Andrews pattern holds, then RDDs will underestimate the effect for our population.
- No pre-registration
- Short followup, during the noisy income ages (the 20s)
- Income data is from 2007-9, i.e. the Recession. They do some sensitivity checks which pass.
- Income data omits Texans who move out of state. We should expect this to bias the results downward (correcting somewhat for the rampant ability bias).

Mountjoy & Hickman (2021)

We can check the above with a good matched-applicant study from Texas using similar data. Uniquely, this gives us a per-college measure: the effect of attending college x instead of the most prestigious college, UT-Austin.

- *Population*: Texan freshmen, 1999-2008
- *Intervention*: Attending UT Austin.
    - *College quality measure*: average intake SAT
- *Control*: Admitted to UT Austin but attended less elite
- *Outcome*: earnings
- *Time*: 8-10 year followup
- *Sample*: n = 422,949; identifying subsample of n=125,867 across 3,283 matching portfolios. Whole-population administrative data. Filtered to applicants applying to >1 college (methodological reasons) and to students and workers who stayed in-state.
- *Matching*: by the set of institutions an applicant was admitted to

This study is great in a few ways:

- Their identification analysis allows for unrestricted heterogeneity in the value-added of each college across different students.
- Their implementation of matching-applicants seems to address selection bias.
    - They show that usual pre-college covariates are not predictive enough to adjust for selection bias.
    - While adjusting for fixed effects of “portfolio of college admissions” nearly eliminates differences in predicted outcomes across colleges, in other worlds, students who apply and are admitted to the same set of institutions are very similar no matter where they actually enroll.
- They explore subtle sorting on the basis of student-specific match effects.
    - They relax the common (but strong) assumption of homogeneous treatment effects. They difference mean student outcomes, across colleges, within admission portfolios.

Result

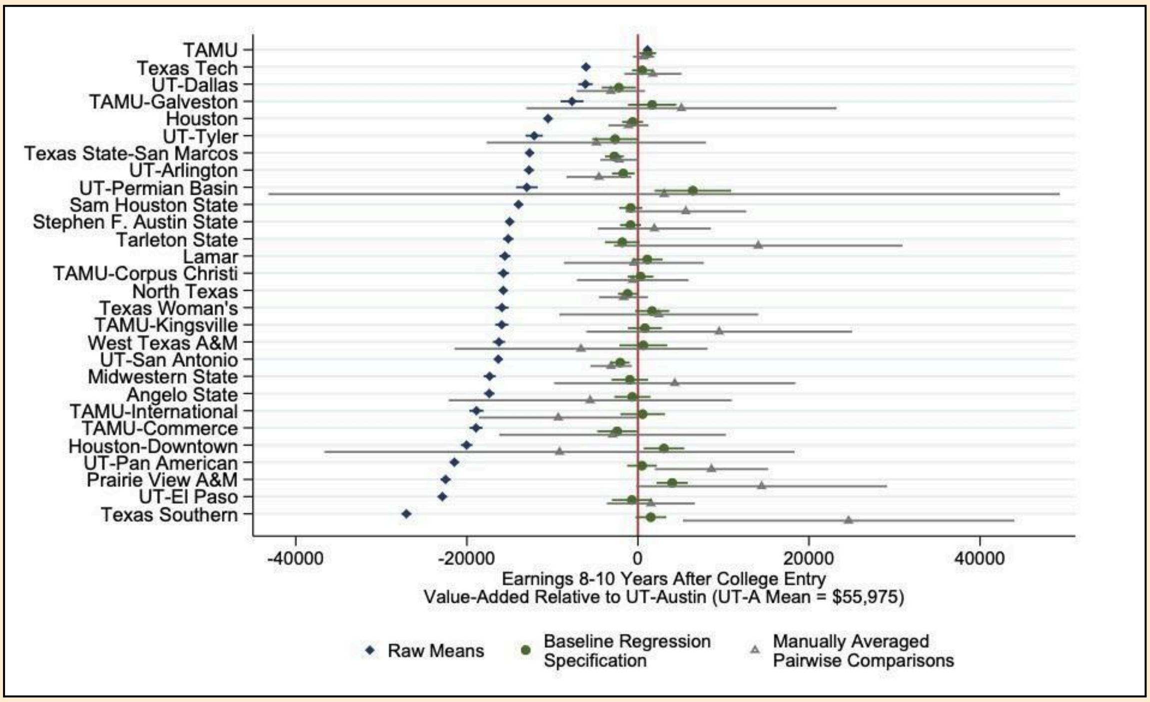


Figure 4. Triangles are the direct comparison with UT-Austin

- Basically no effect of +100 average intake SAT on earnings
- From eyeballing the figure 4, there is no or at most small effect of not going to UT Austin (a much better school, world top 100), x10 less than one would have naively guessed from average salaries alone.
- Dozens of specifications are tested, and some find something. e.g. Figure A.7 (bottom) uses the modal college as the comparator to the top college (Austin) and finds ~$4k higher earnings for UT Austin (about +9% income).
- **Value-added varies a lot across colleges and while selectivity is a nearly perfect predictor of raw outcome means, it is a poor predictor of value-added within student choice sets.**

Figure 6: Predicting Raw Mean Earnings vs. Value-Added with College Selectivity

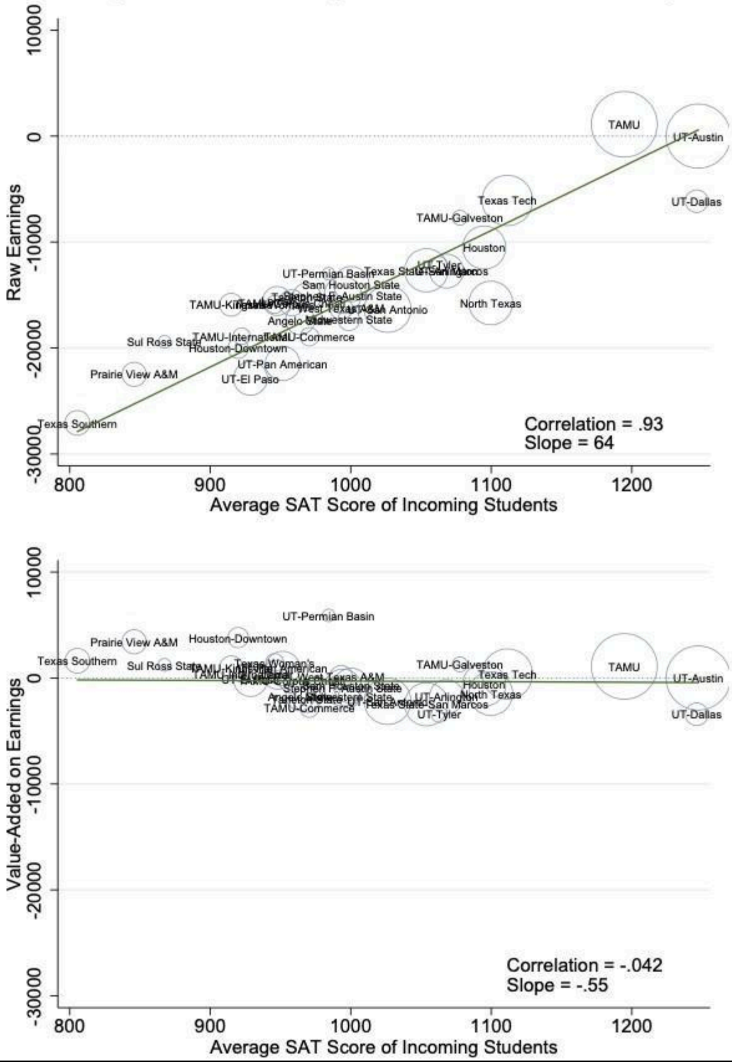
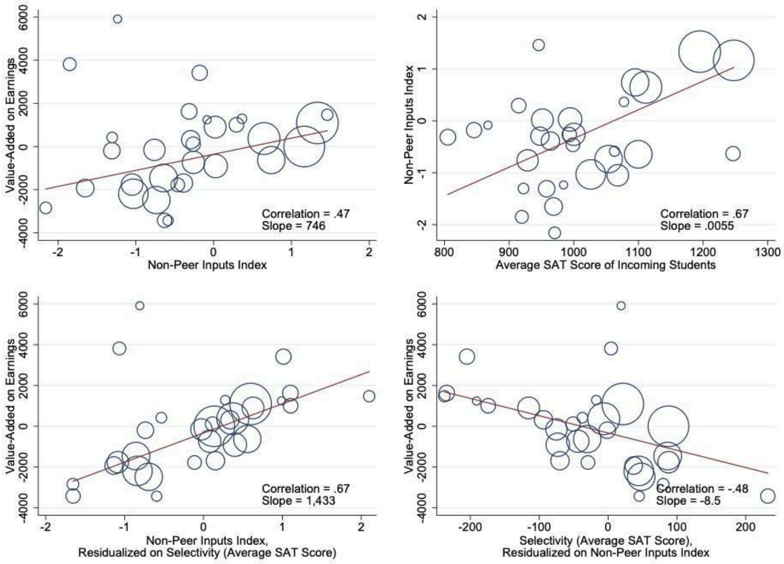


Figure 10: Peer vs. Non-Peer Inputs and Value-Added



Limits

- Just Texas. UT-Austin is in the top 40 in the US; *maybe* there's some difference above that.
- Matching is still observational (but it seems better than any extended set of covariates).
- They intentionally squashed extreme outcomes ("We winsorize at the top and bottom 0.1th percentiles").
- One problem with using administrative income data is that it doesn't include the 20% of income rich people manage to avoid tax on. It also seems to ignore capital gains income.
- No pre-registration.
- Didn't catch Texans who went out of state. But "*Texas has the lowest outmigration rate of any state in the U.S… less than 4 percent*". But this will include the top students.

Dobbie & Fryer (2014)

Since the evidence about top percentile students is scarce, we include an RDD about elite high schools. This is a clean study including corrections for multiple comparisons and continuity tests, with a hard endpoint (properties of the college they later attend).

- *Population*: 4 high schools in NY, average SAT 99th to 99.9th percentile
- *Intervention*: RDD. Being admitted to an extremely selective high school
    -
- *School quality measure*: top average intake SAT
- *Control*: applicants whose exam score was just below the cutoff
- *Outcome*: four-year enrollment; four-year graduation; enrollment into a college with high average SAT; postgrad enrollment
- *Time*: 3 to 7 year followup
- *Sample*: n = c. 39,000

Result

Applicants just eligible for an exam school have *peers* that score +d=0.17 to 0.36 higher on 8th-grade state tests.

However, exposure to these higher-achieving (and more homogeneous) peers has little impact on the cutoff students' college enrollment, college graduation, or college quality. (Maybe even slightly worsen these outcomes but no statistical significance if we adjust for multiple comparisons).

Limits

- *RDDs measure the worst of the best*. This is less relevant here since the cutoff is so high.
    - But still the effect of programmes on *marginal* admitted students could be different from the effects on students in the opposite tail. (This might matter because sometimes the limiting factor is financial aid, and not test scores.)
- No pre-registration
- Not really long run outcomes
- It's unclear if students get more or less out of college:
    - e.g., someone was gifted in STEM but discovered theater at school and enrolled into movie school; school has lower avg SAT but person is following his true passion
    - e.g., someone would have went to the most prestigious university possible but decided to go to another one with a strong lab doing work related to his Intel Science Fair project
    - cf. Estrada and Gignoux (2017) showed that graduates of National Polytechnic Institute's high schools enjoy an additional 15% increase in income (on top of 100% college premium).
        - But note that literature suggests that more selective high schools

tend to produce all sorts of benefits (like learning gains) in some low/middle-income countries, while no such gains are observed in the developed countries.

- A 2011 paper by the same authors checked that variables are changing smoothly around the cutoff, so their identification assumption is checked.
- Some (limited, noncausal) counterpoints: Elison and Swanson (2016) find that schools have a strong and persistent (across years) effect on students' performance at American Mathematics Competitions, specifically at the very far right tail (>99.9% math ability).
    - It's unclear to us how much success at AMC and developing corresponding math talent translates into real life outcomes, as opposed to being a mixture of selection for innate talent and careful test preparation.
    - At the same time, Agarwal and Gaule (2018) observed that International Math Olympiad gold medalists are 50 times more likely to become Fields medalists than a PhD graduate from a top-10 math program.