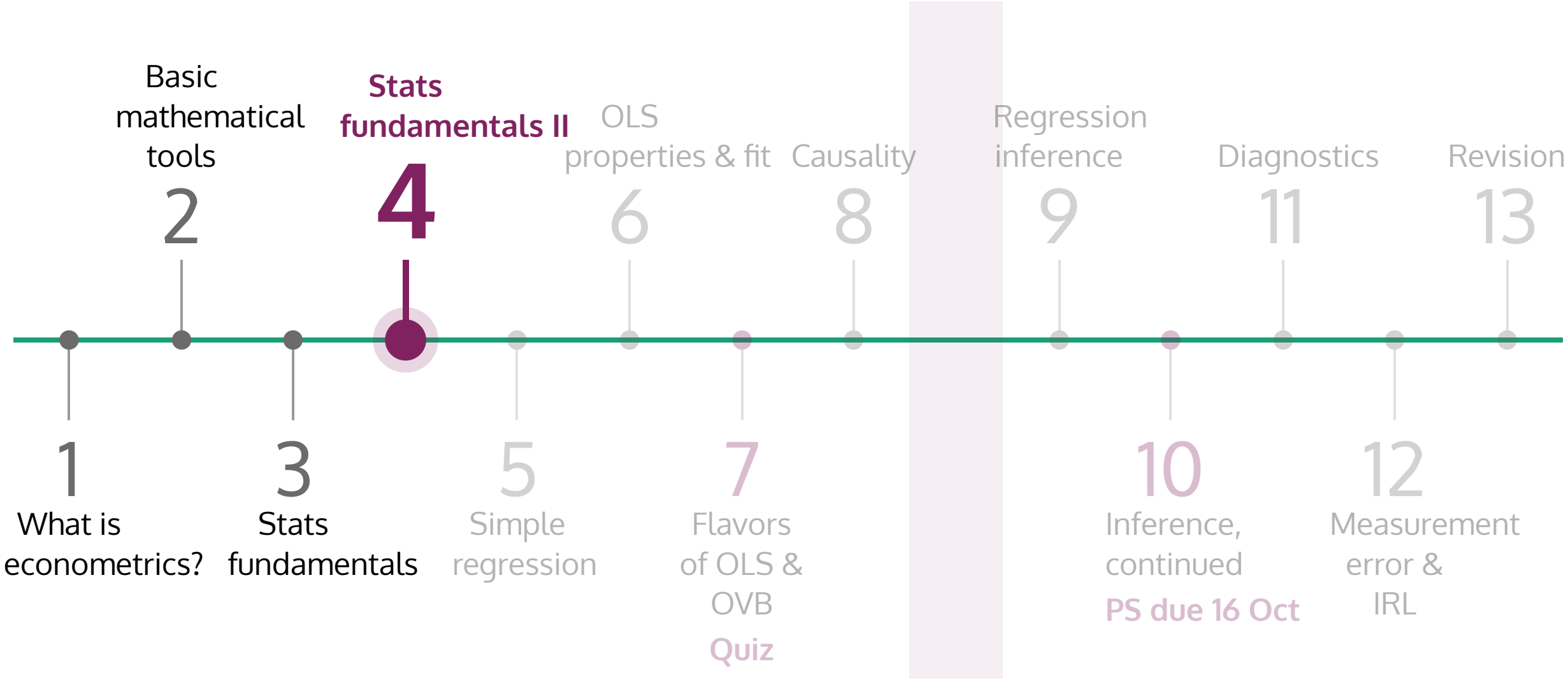

ECON 2041
Introductory Econometrics
Great expectations

Plan for the semester



Plan for today

- Your grade forecasts, revealed
- Income and life expectancy in the US
- Discrete or categorical?
- A correlation gallery
- Where a covariance comes from

What you expect of yourself

What I asked you

Your own semester

- Expected final mark (0-100)
- Plans:
  - attend live?
  - study hours?
  - try tutorials before class?
- The mark if you did "everything"

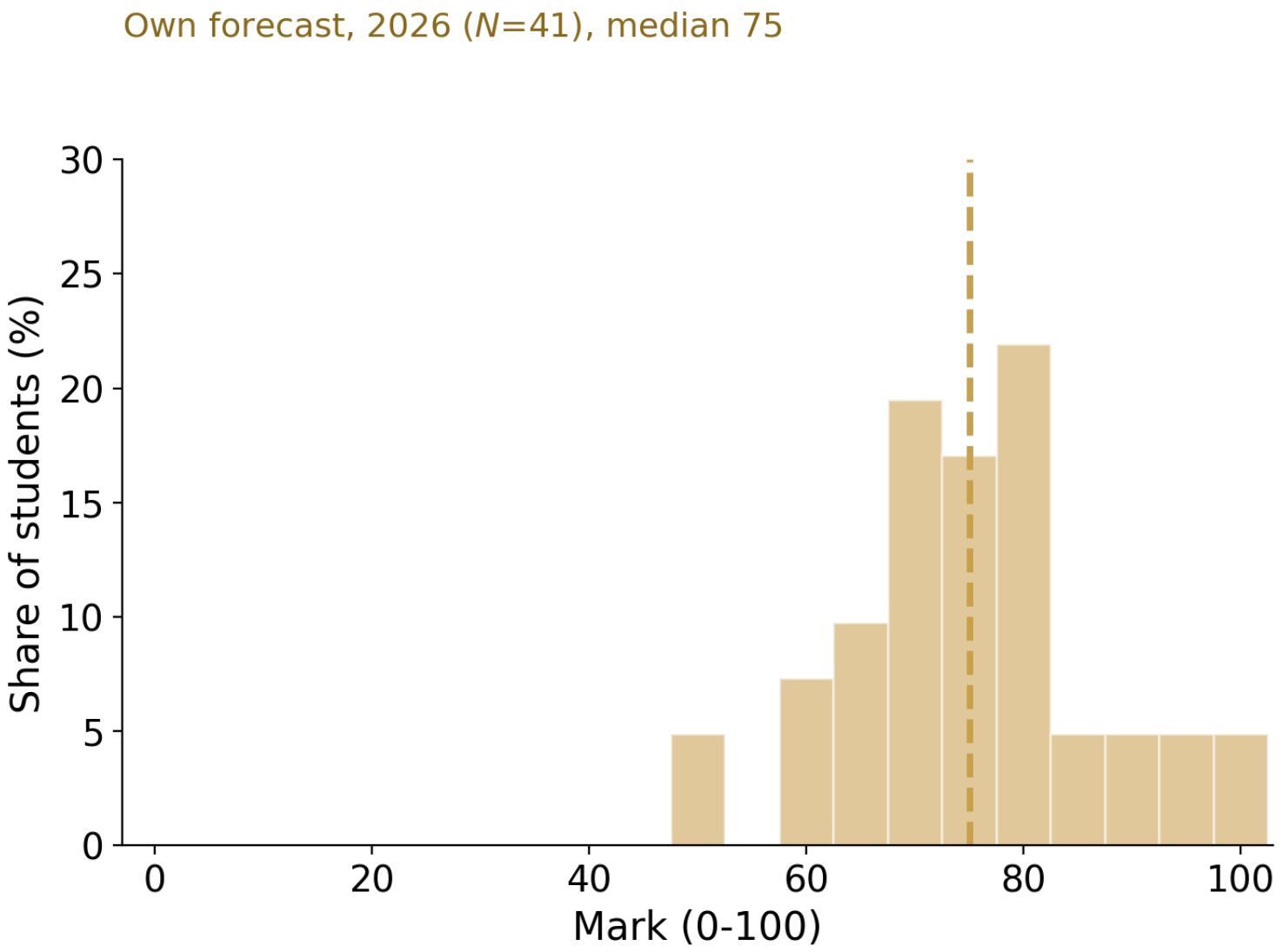
Three hypothetical students

- **Attends live lecture:** rarely / some weeks / most
- **Study hours:** <2 / 2-5 / 5+
- **Try tutorials first:** no / yes

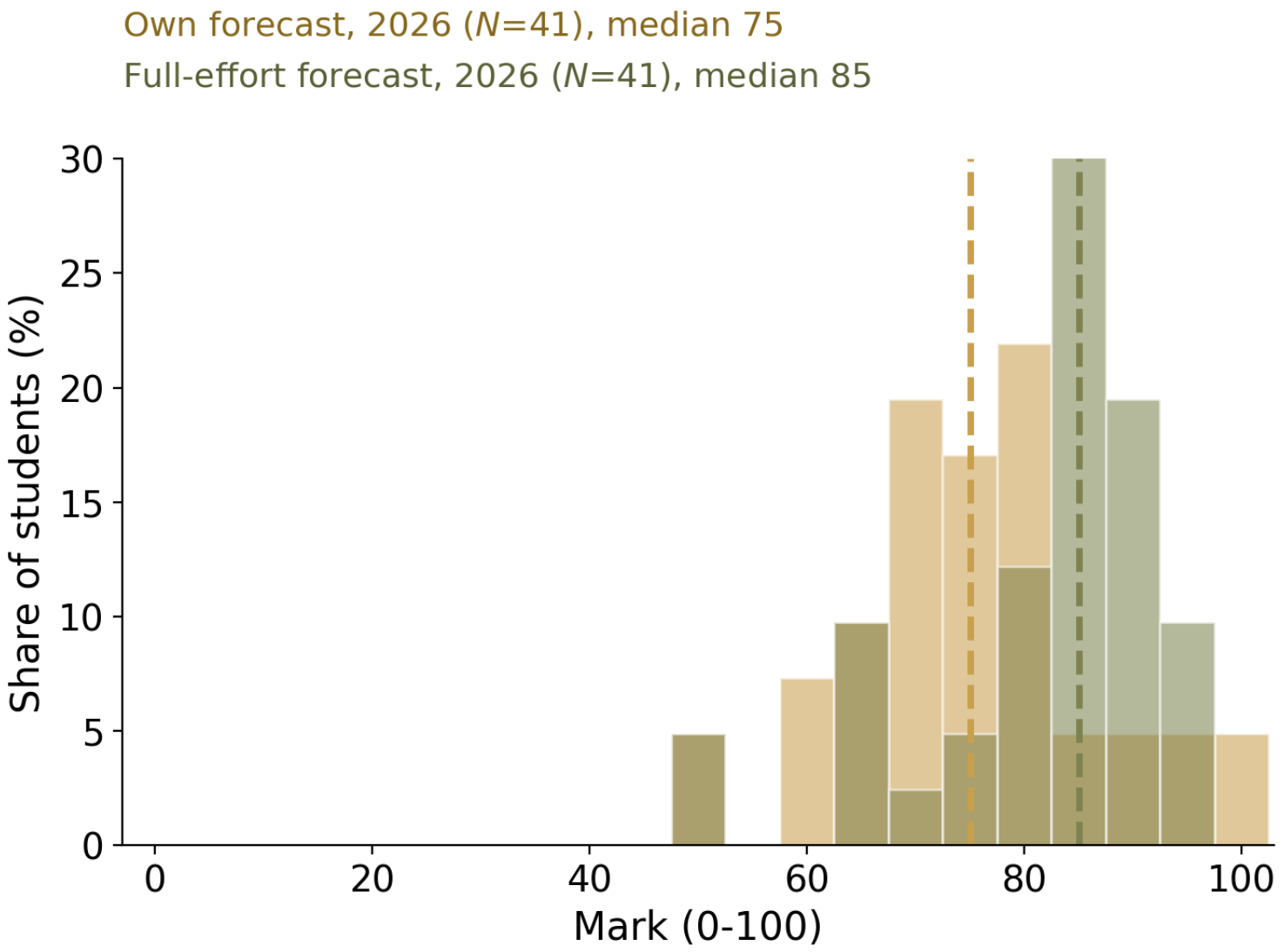
For each:

- a mark
- % chance they pass,
- % chance they get a D

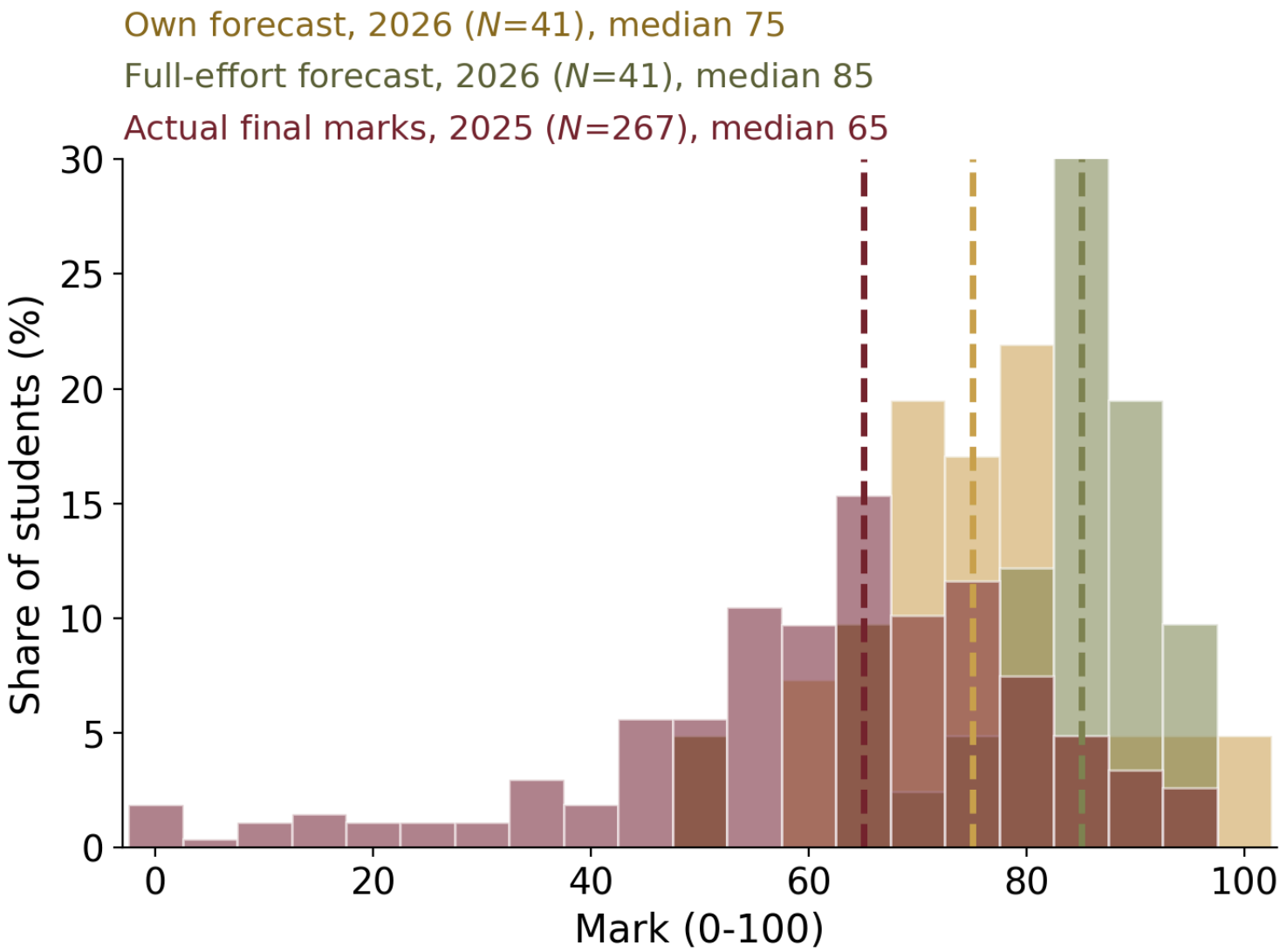
Nobody expects to fail



Full effort: you expect it to add about 7 points!

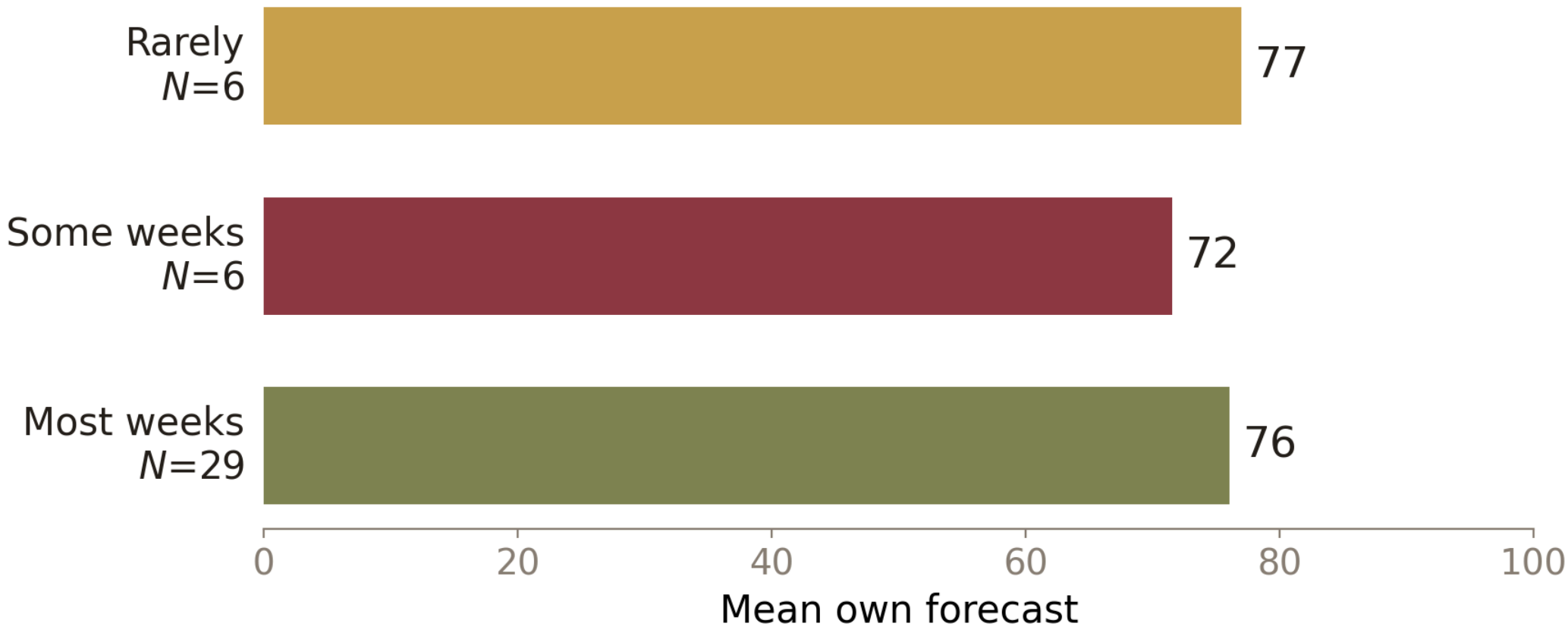


Last year's marks

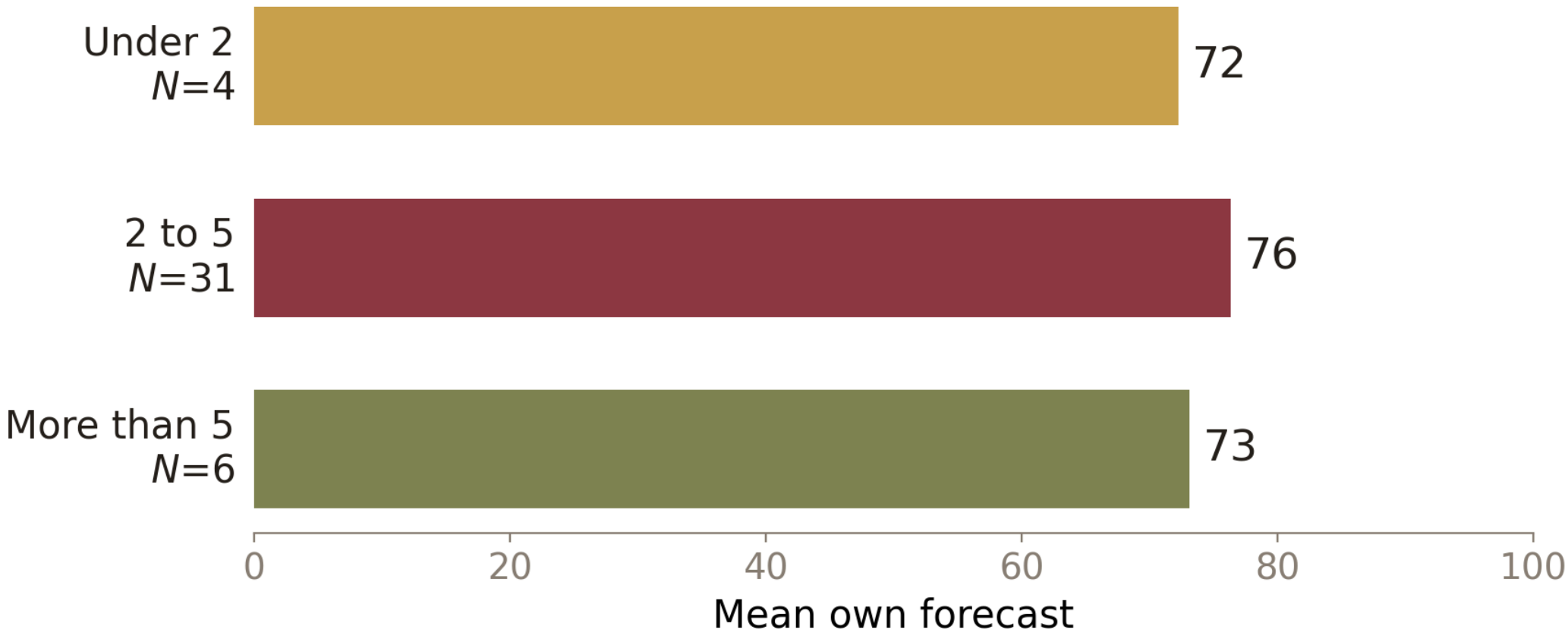


Your forecast, given your plans

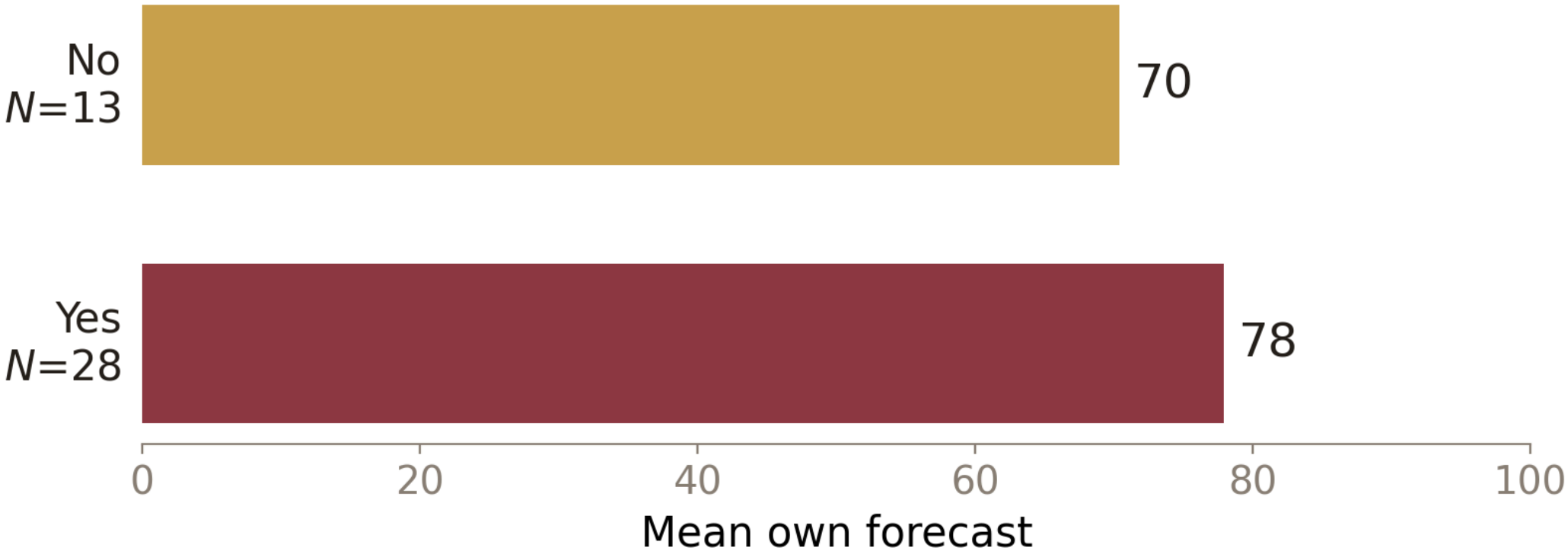
Attendance plans: flat



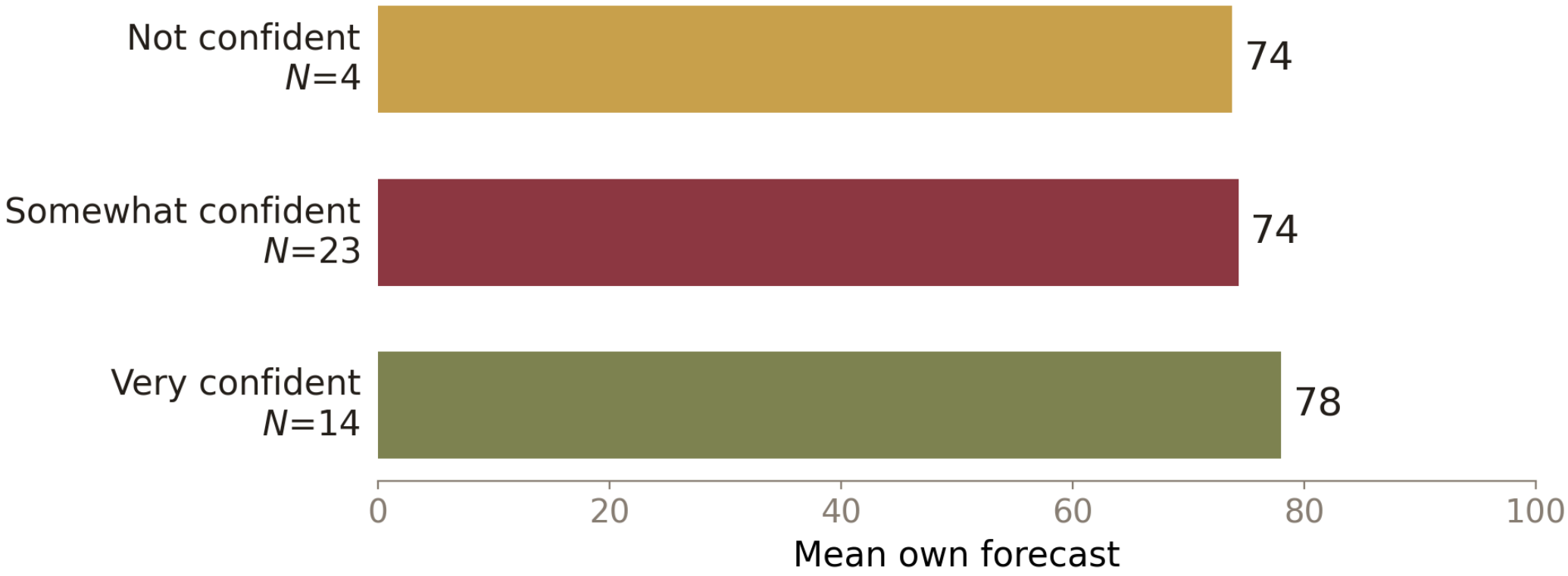
Study-hour plans: flat



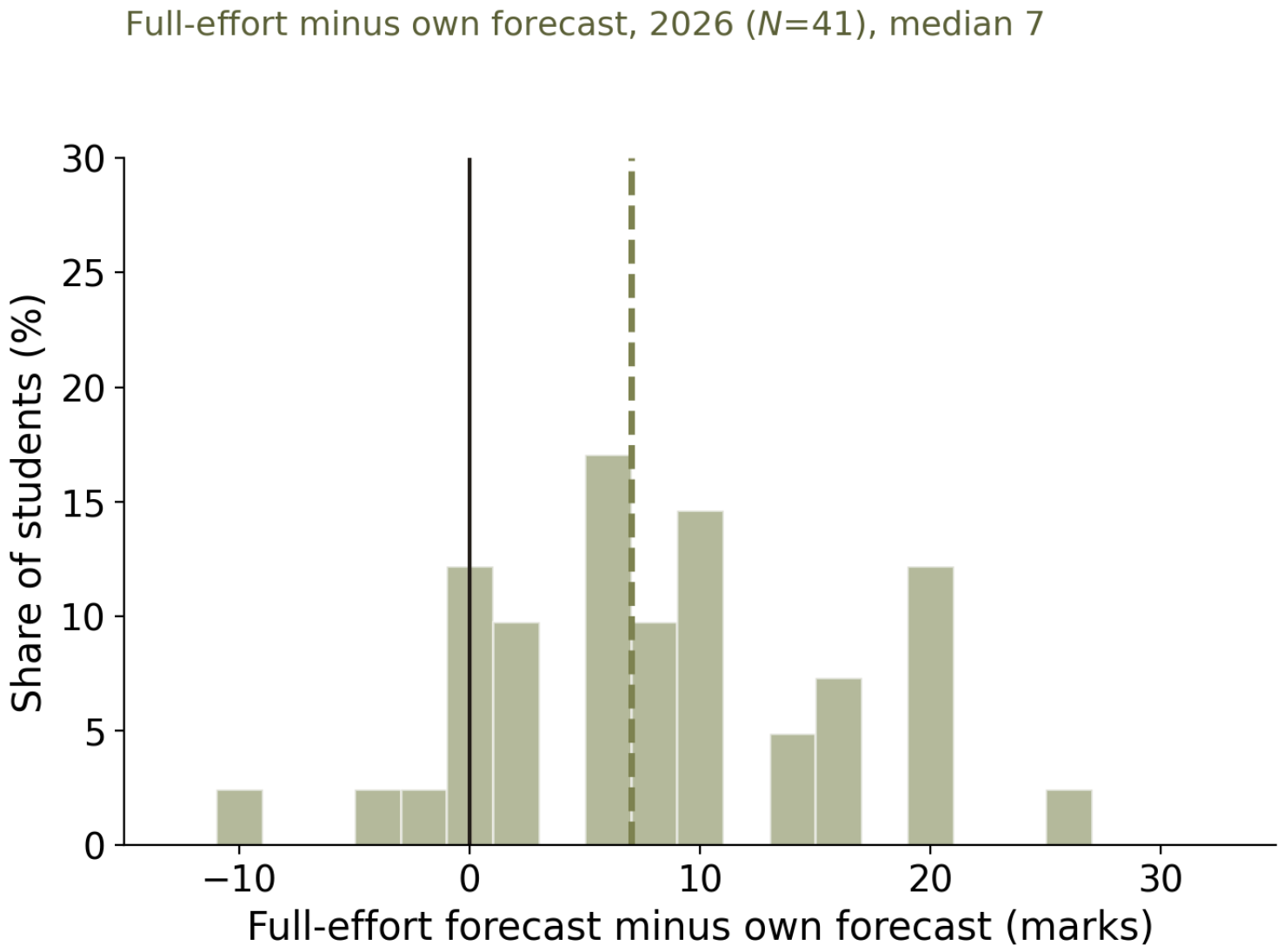
Tutorials before class: where the action is



Confidence in the plan barely matters

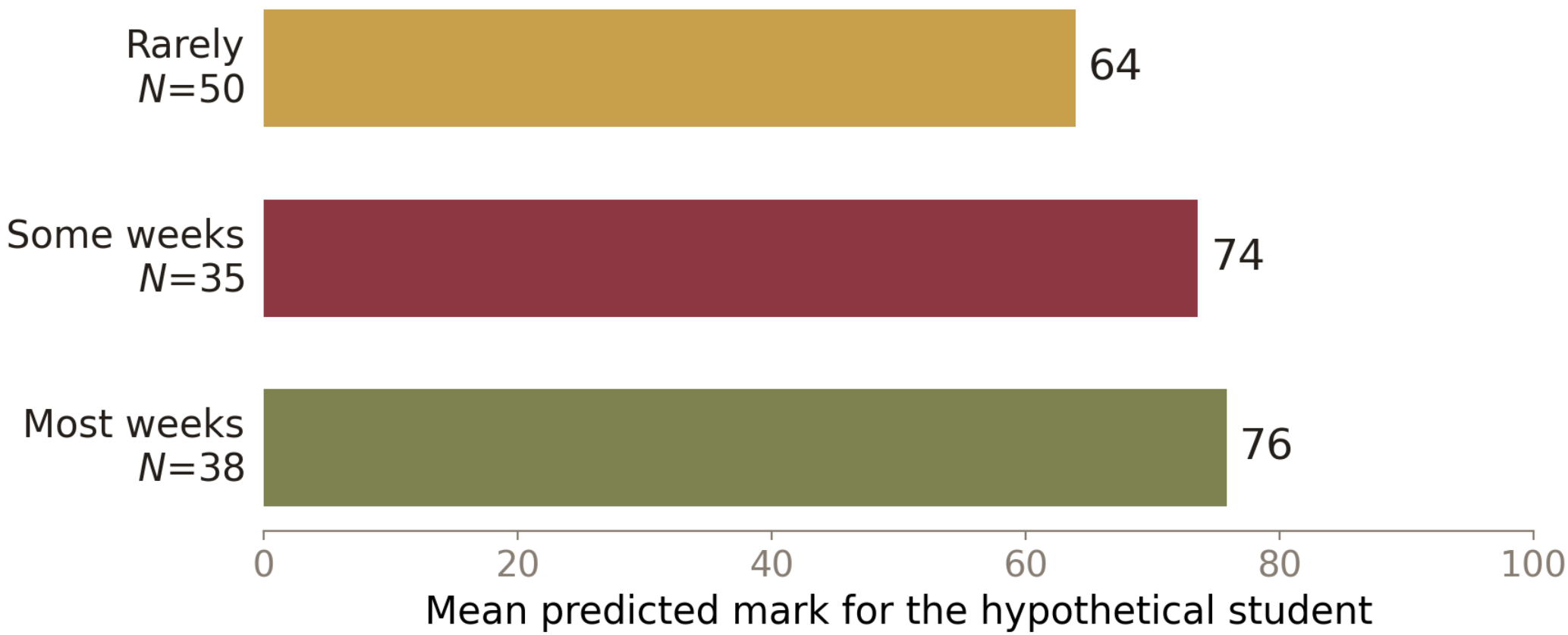


Effort, student by student

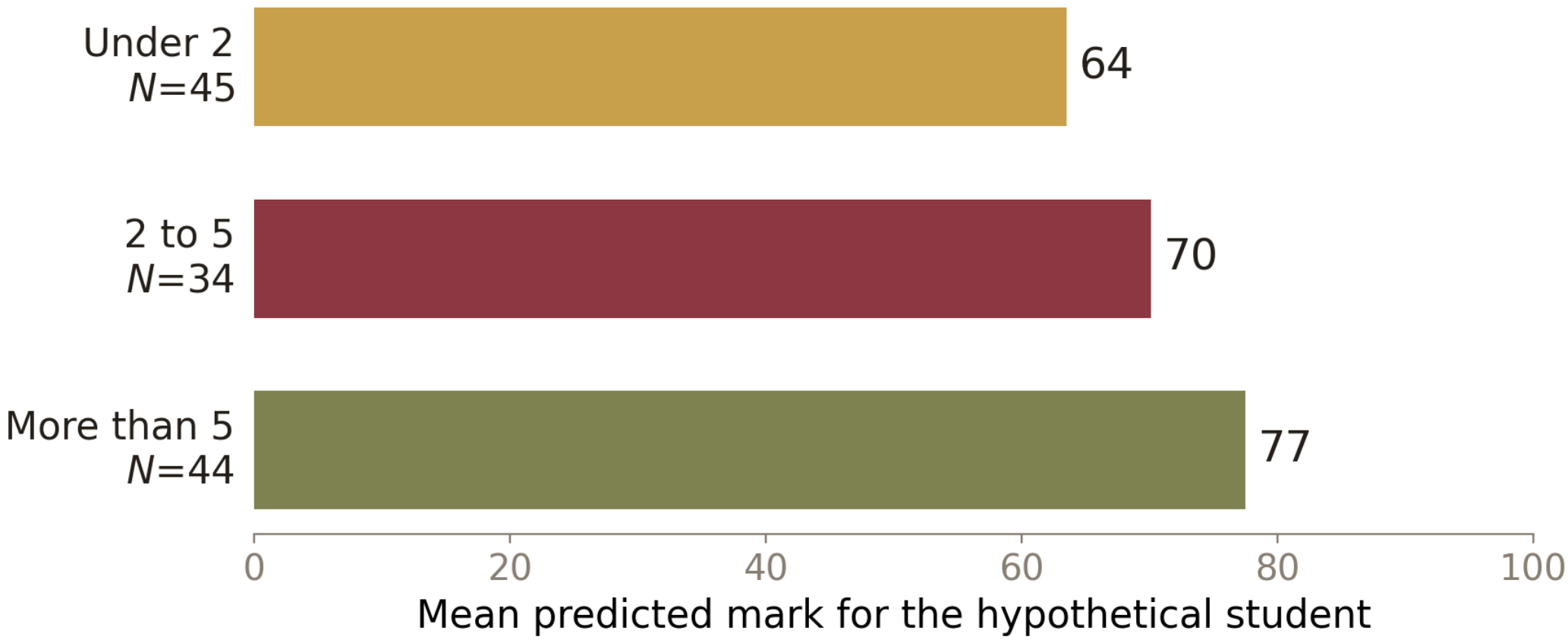


What you expect of others

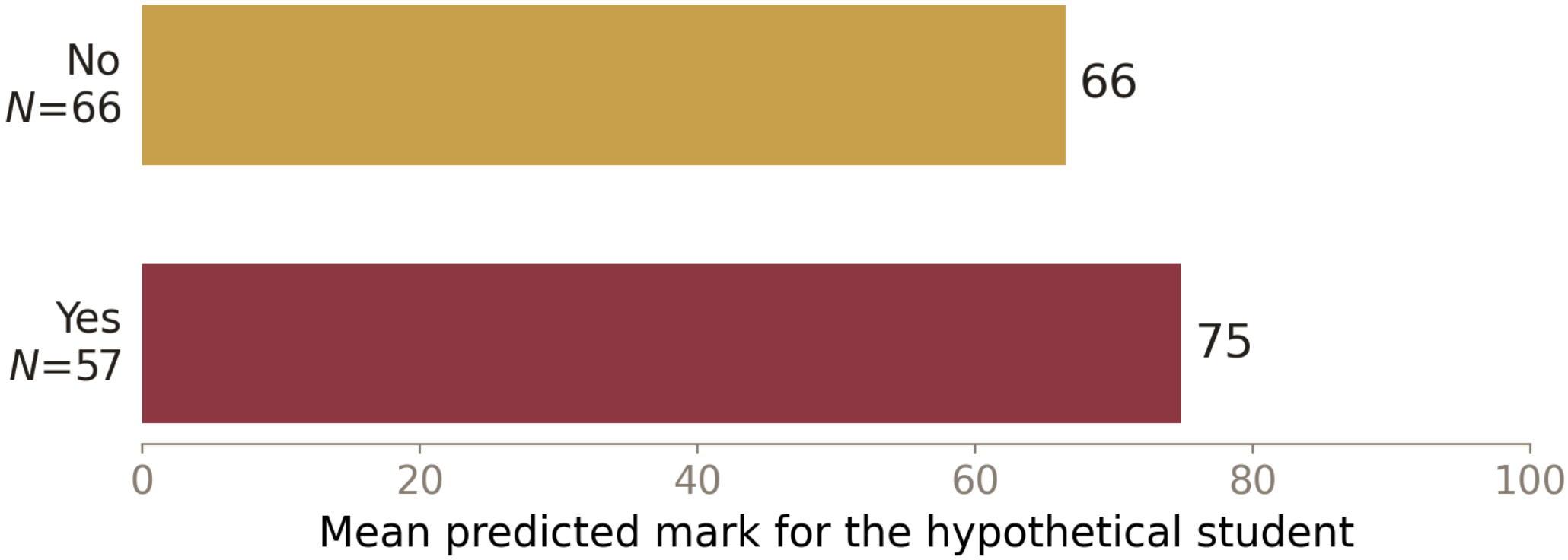
You DO think that attendance matters



You also think that study hours matter



Same story for tutorial work before class

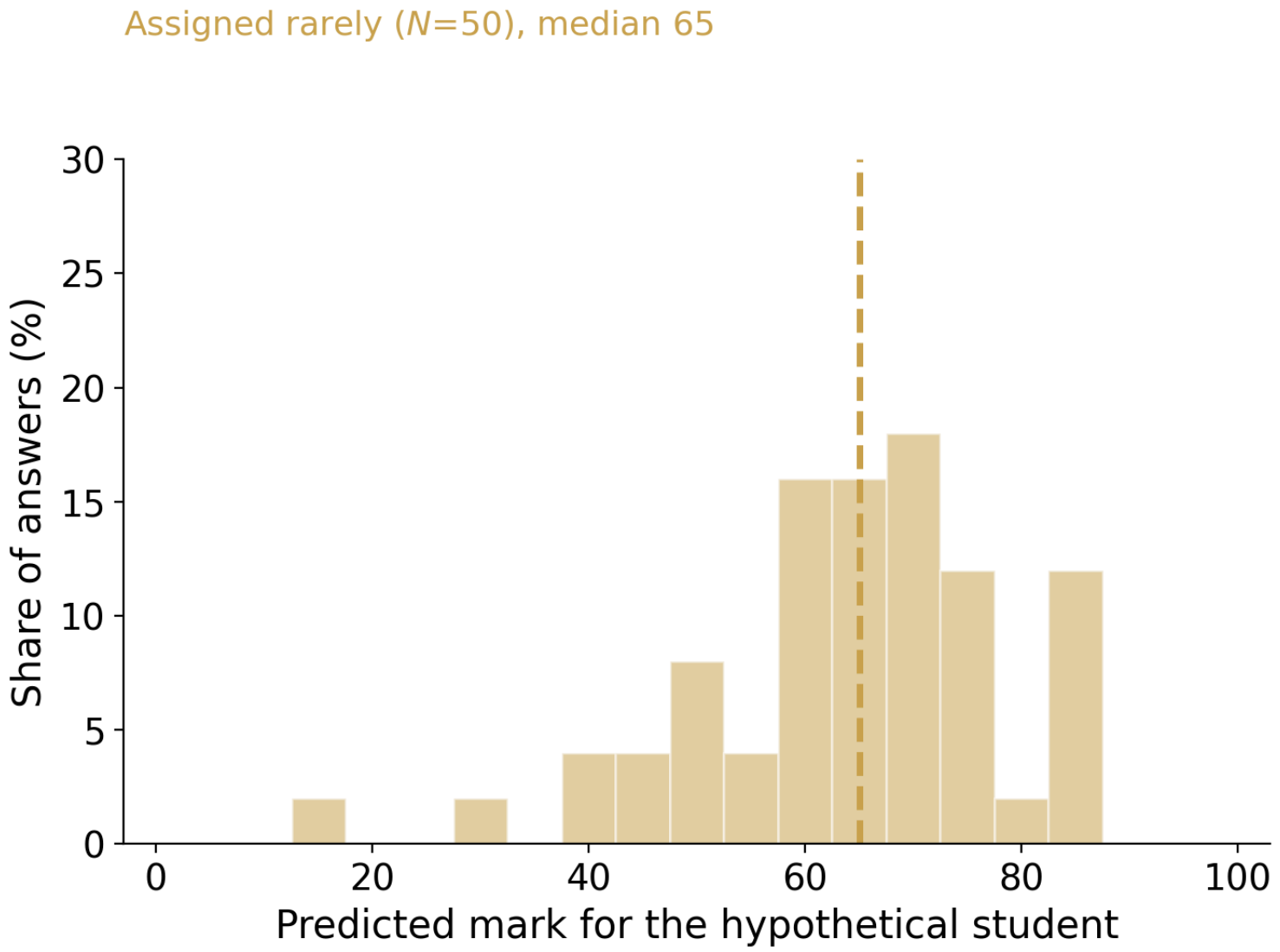


Independence by design

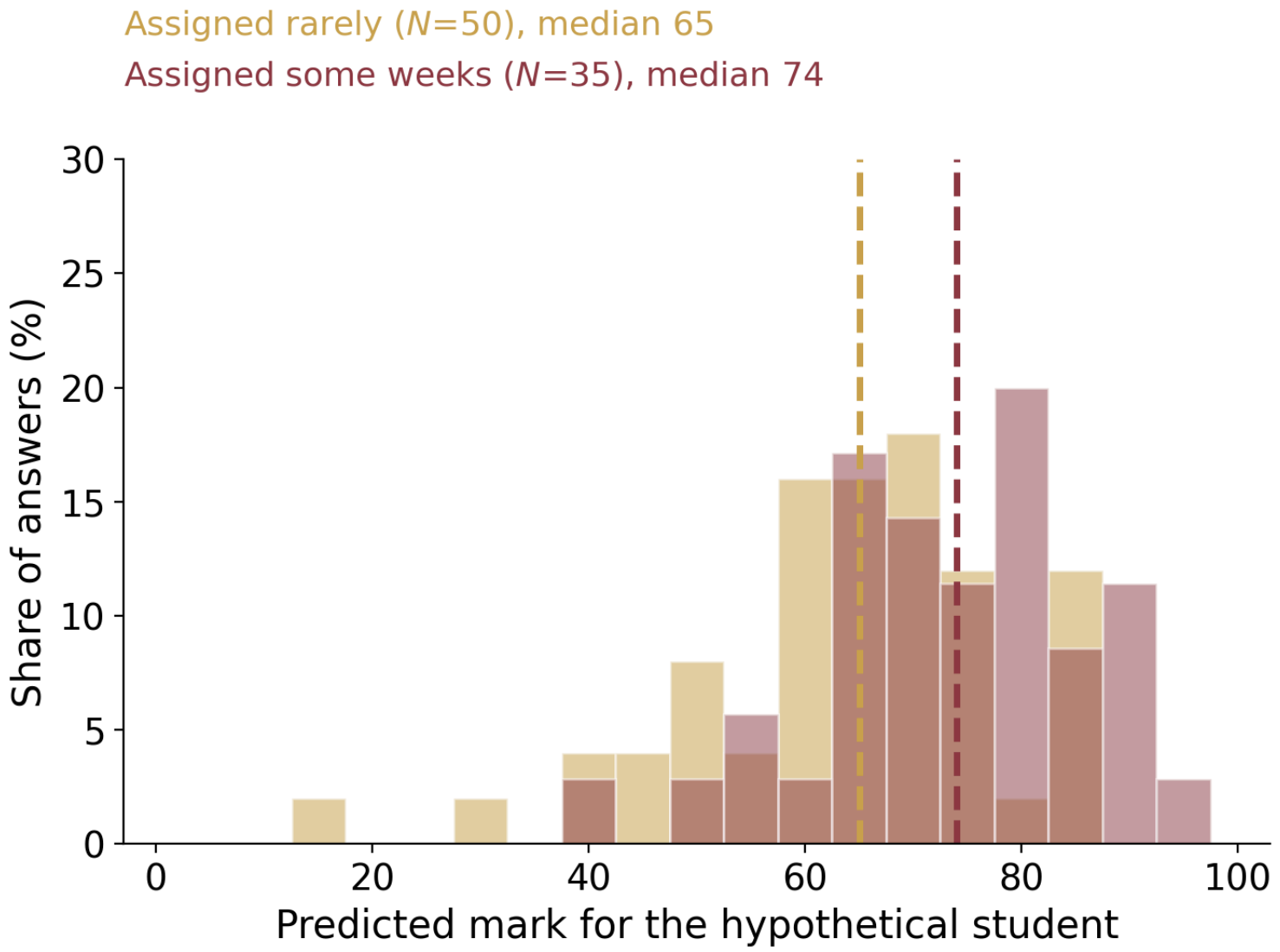
Share of hypothetical students at each study-hours level, within each attendance level

	Assigned study hours per week			
Assigned attendance	**<2**	**2 - 5**	**5+**	***N***
Rarely	44%	30%	26%	50
Some weeks	20%	26%	54%	35
Most weeks	42%	26%	32%	38
All vignettes	37%	28%	36%	123

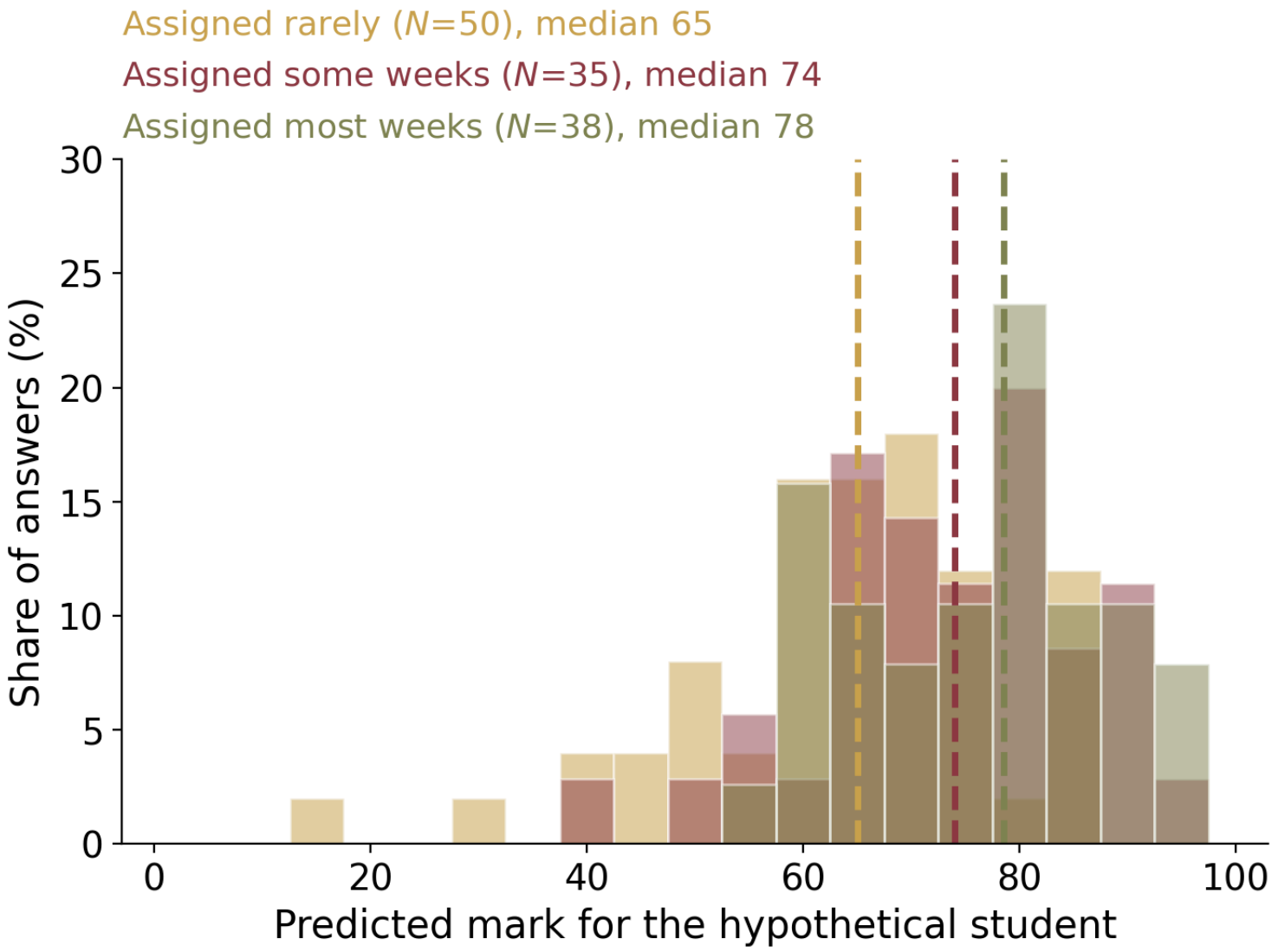
Predicted marks, given "rarely"



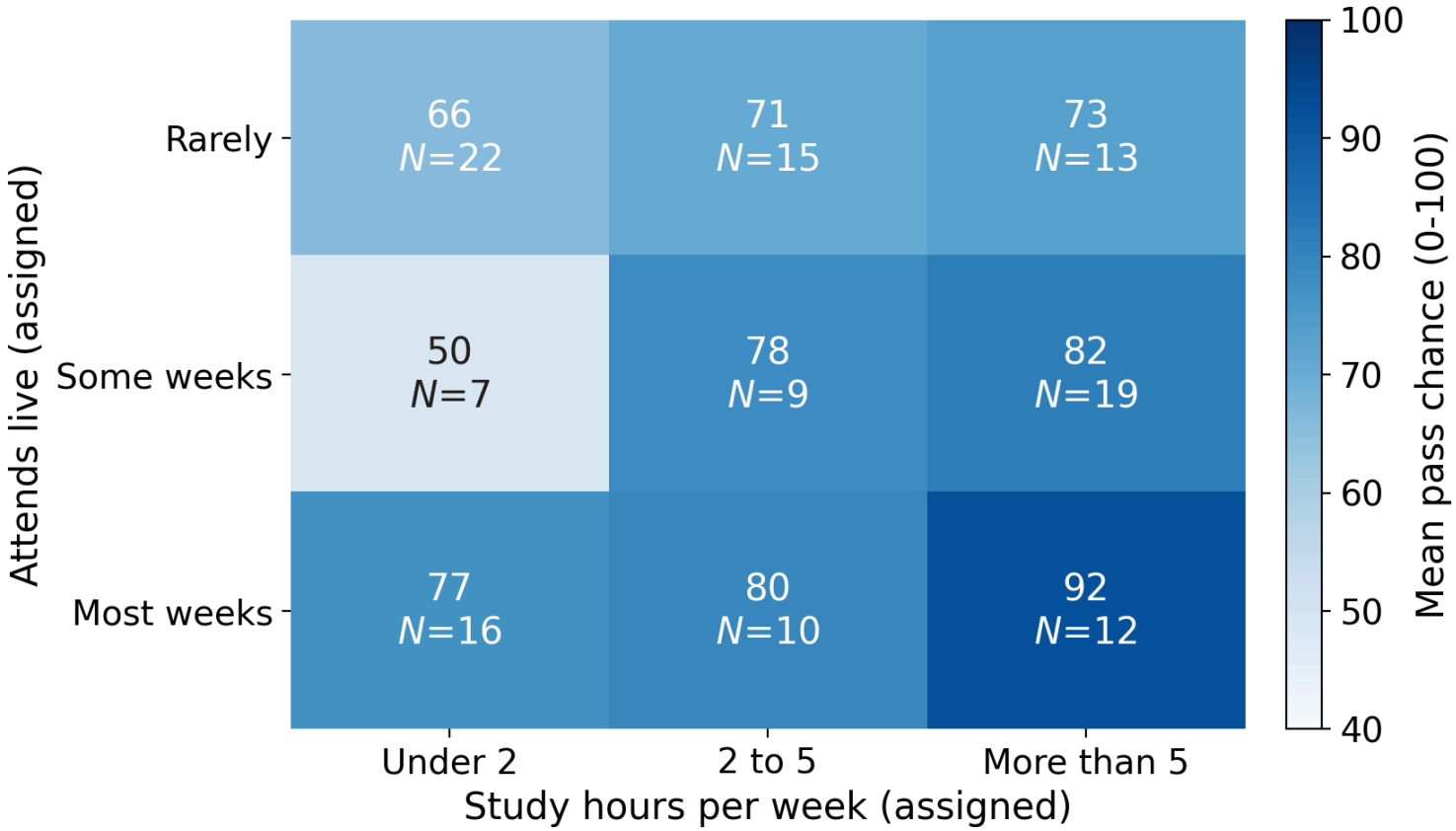
Predicted marks, given "some weeks"



Marks are not independent of attendance



Pass chance, given attendance & hours



Distinction $\Rightarrow$ pass

For each hypothetical student, two chances:

- $P(\text{mark} \geq 50)$, the chance of a pass
- $P(\text{mark} \geq 75)$, the chance of a distinction

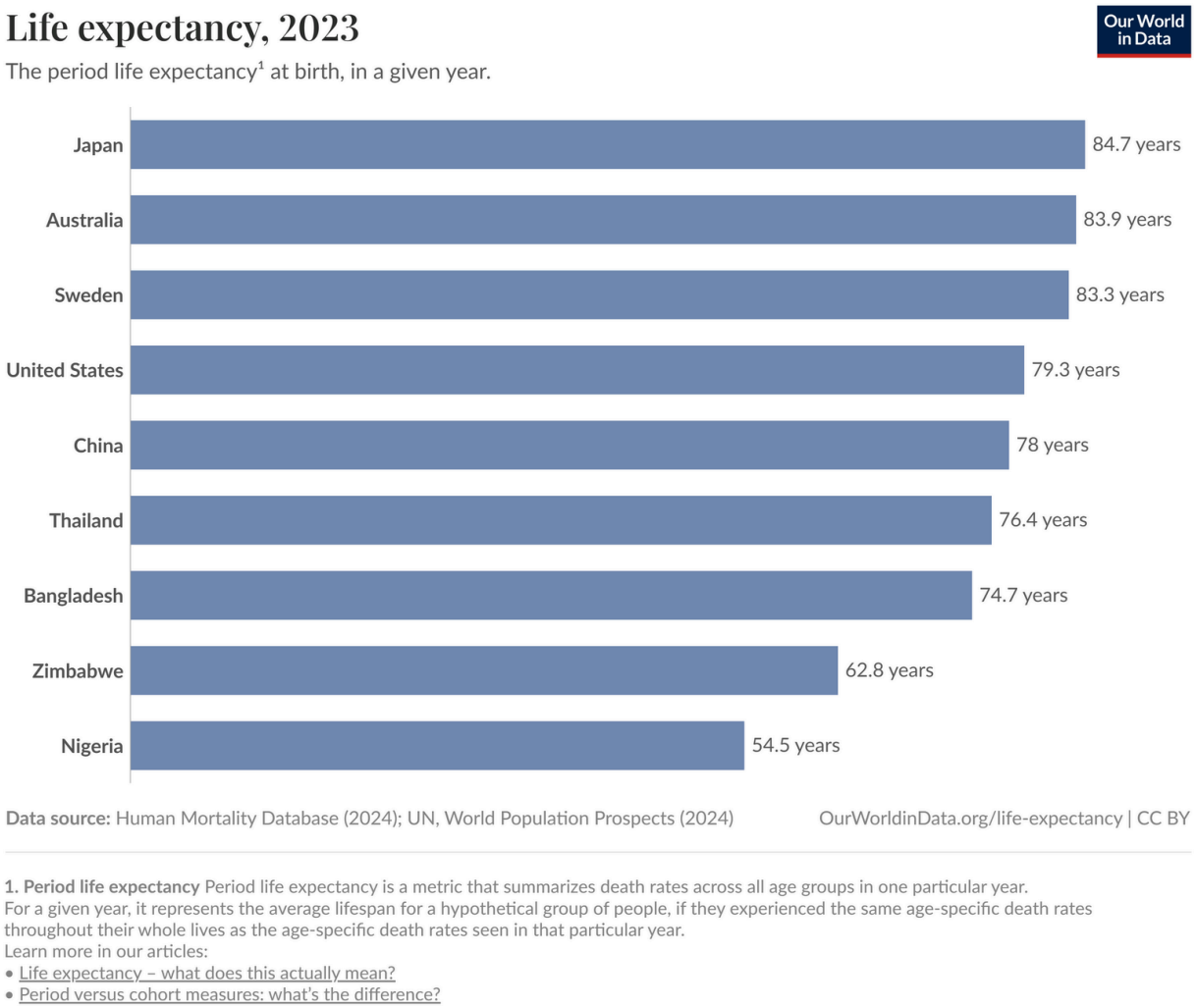
Every mark of 75 or more is also a mark of 50 or more

$\{\text{mark} \geq 75\} \subset \{\text{mark} \geq 50\}$, so $P(\text{mark} \geq 75) \leq P(\text{mark} \geq 50)$

10 of 120 answers put the probability of D > probability of P

Income and life expectancy

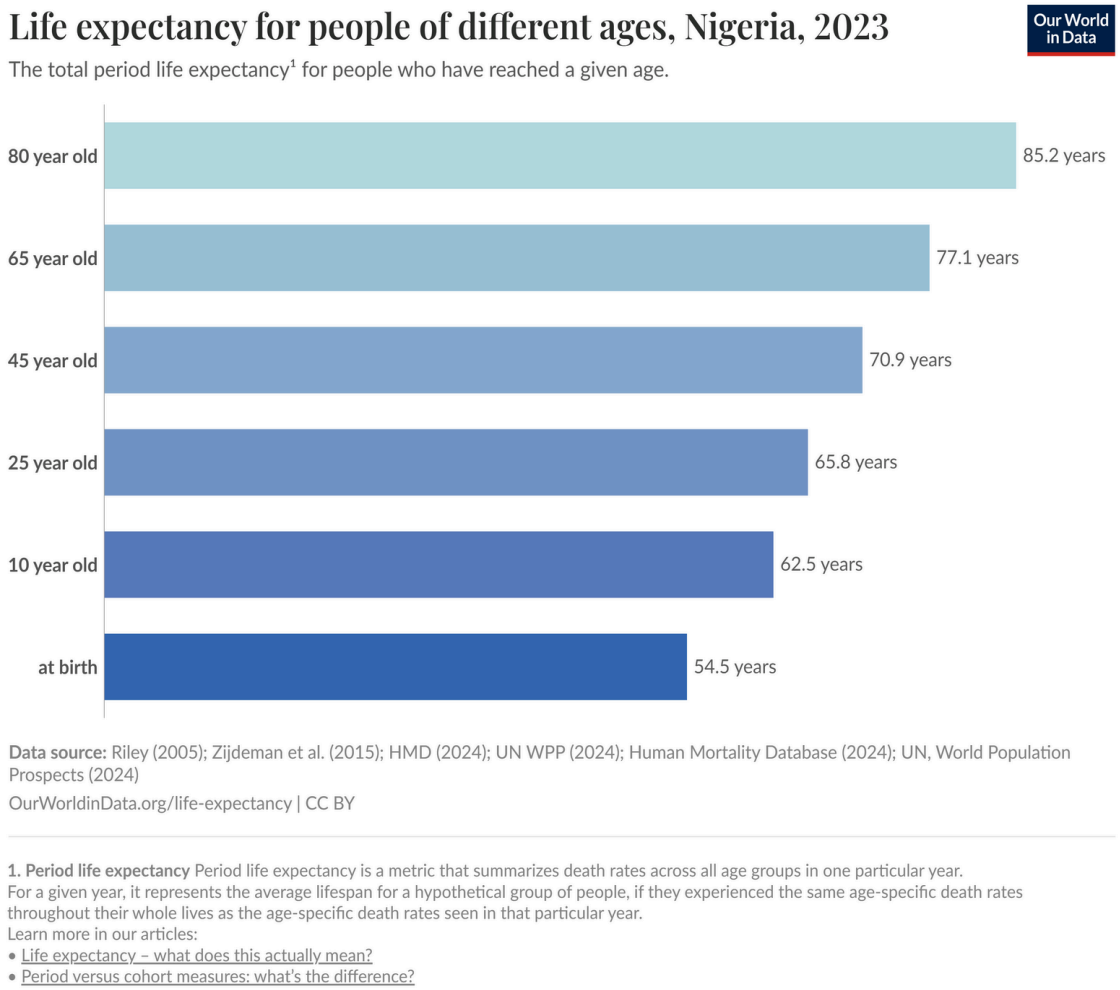
Life expectancy at birth



Life expectancy is based on an estimate of the average age that members of a particular population group will be when they die

Source:
Our World in Data

Life expectancy by current age



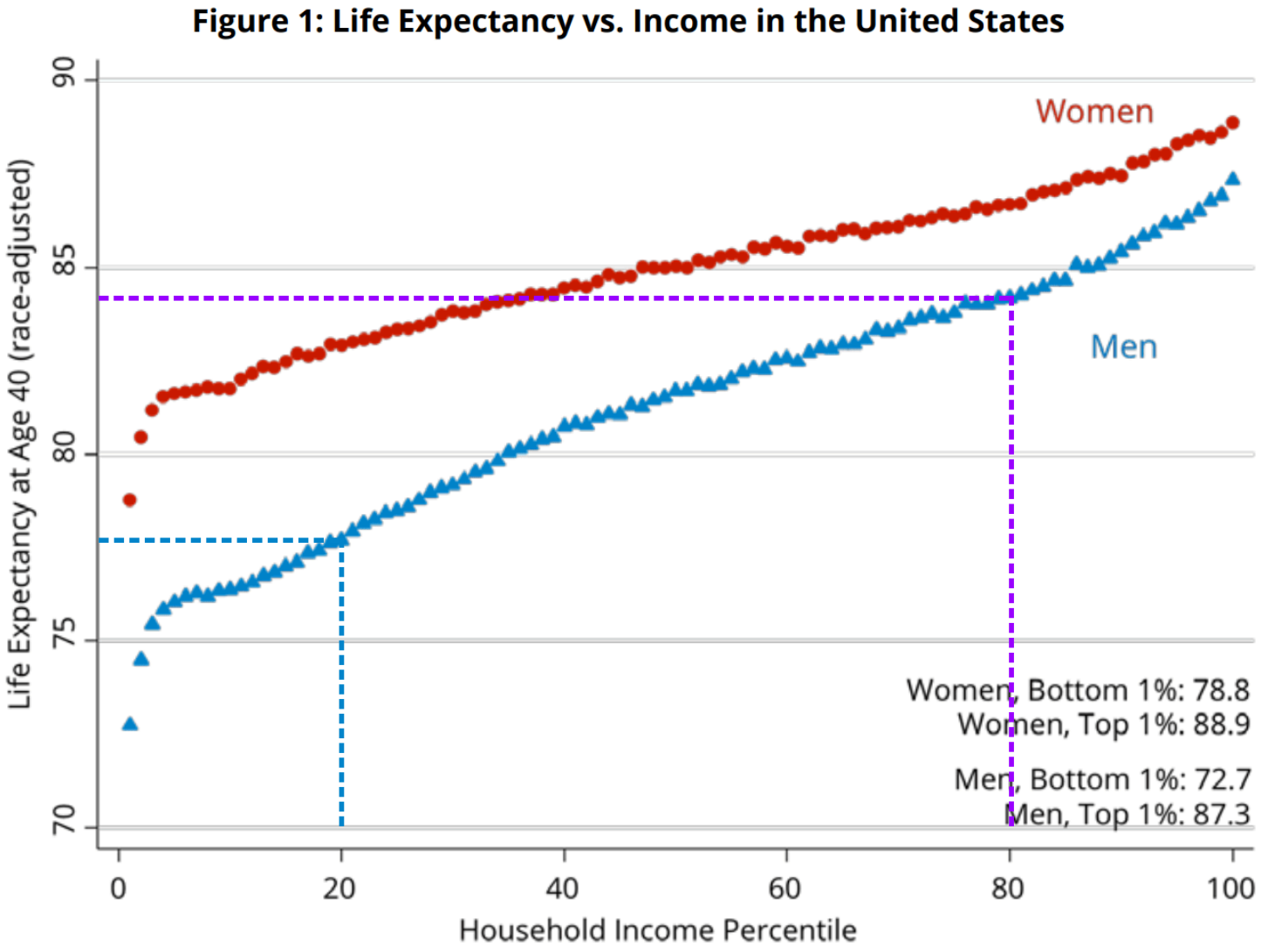
Source: Our World in Data

Guess the gap

US men, aged 40, 2001 - 2014

How many more years does a man in the richest 1% expect to live than a man in the poorest 1%?

The gap is 14.6 years



Men, age 40

- Poorest 1%: 72.7 years
- Richest 1%: 87.3 years

Women: a gap of 10.1 years

Y: life expectancy at 40

X: income percentile

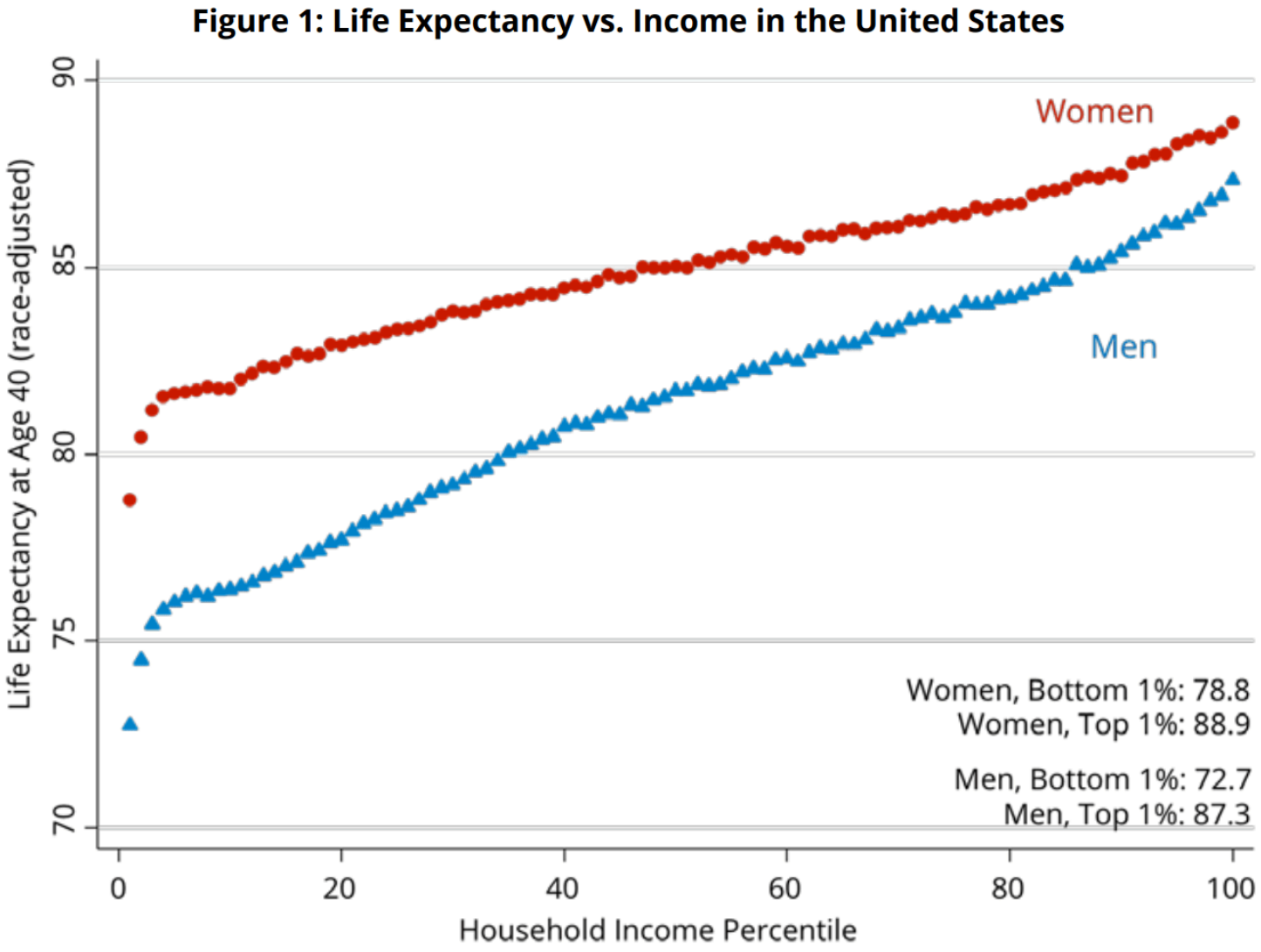
Each dot is $E[Y \mid X = x]$

$E[Y \mid X = \text{p20}] \approx 77$

$E[Y \mid X = \text{p80}] \approx 84$

Source: Chetty et al. (2016)

We also condition on sex here



Y: life expectancy at 40

X: income percentile

Z: sex

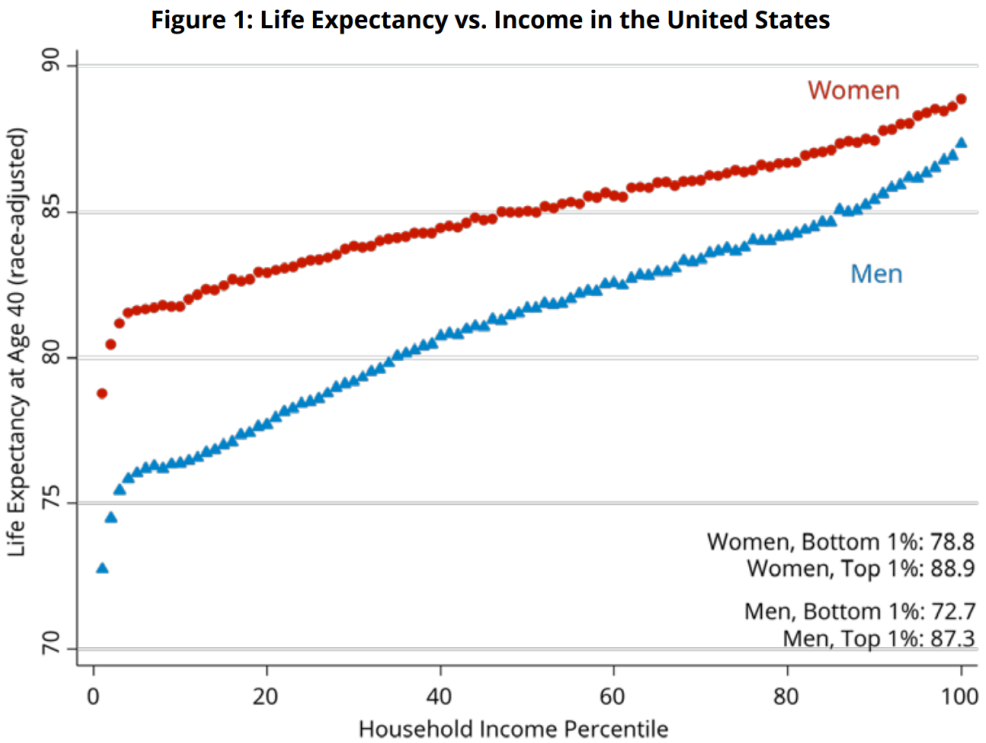
Each line is

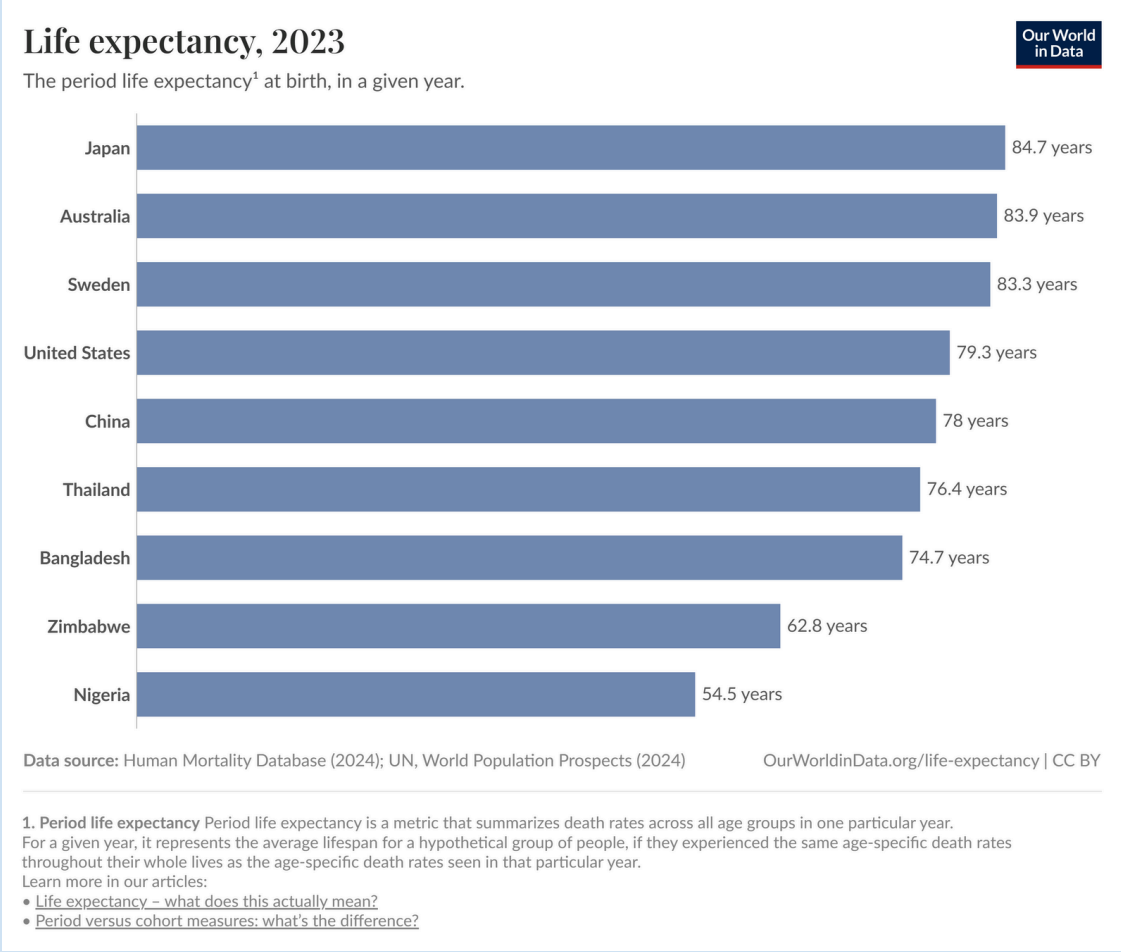
$$E[Y \mid X = x, Z = z]$$

Source: Chetty et al. (2016)

Unconditional vs. conditional expectation

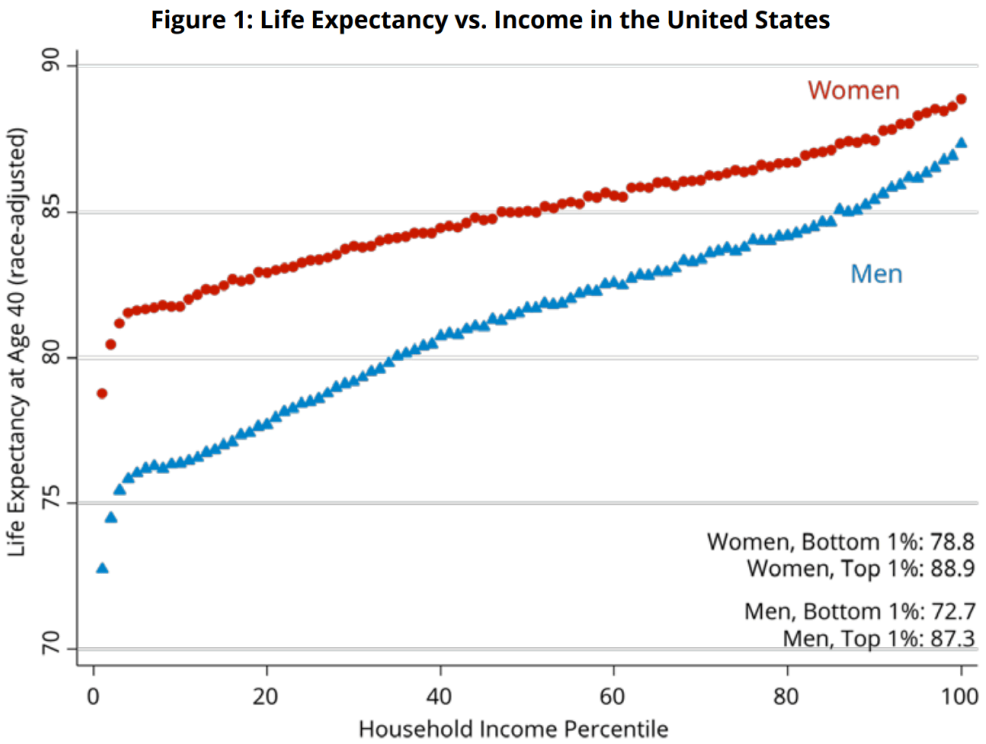
$E[Y]$ hides a lot of heterogeneity that we can see with $E[Y|X]$!

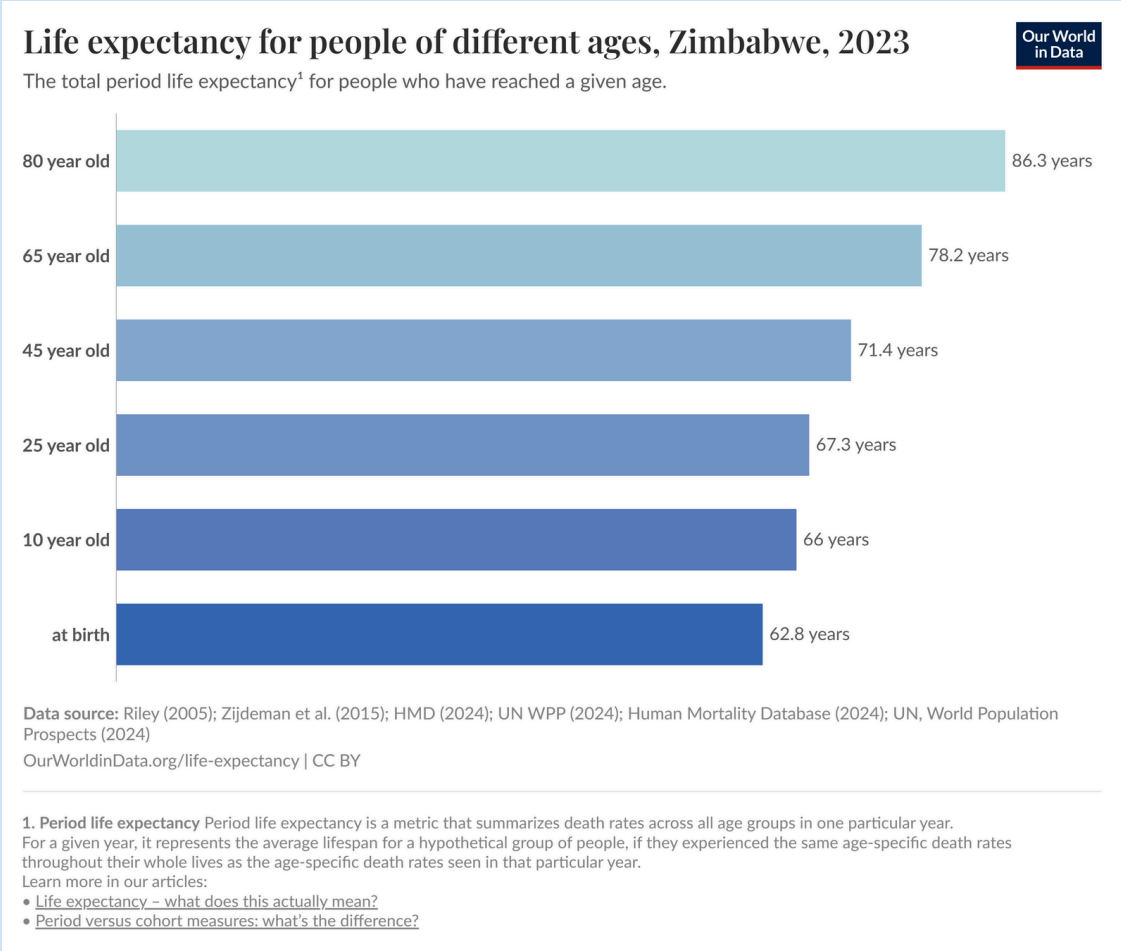




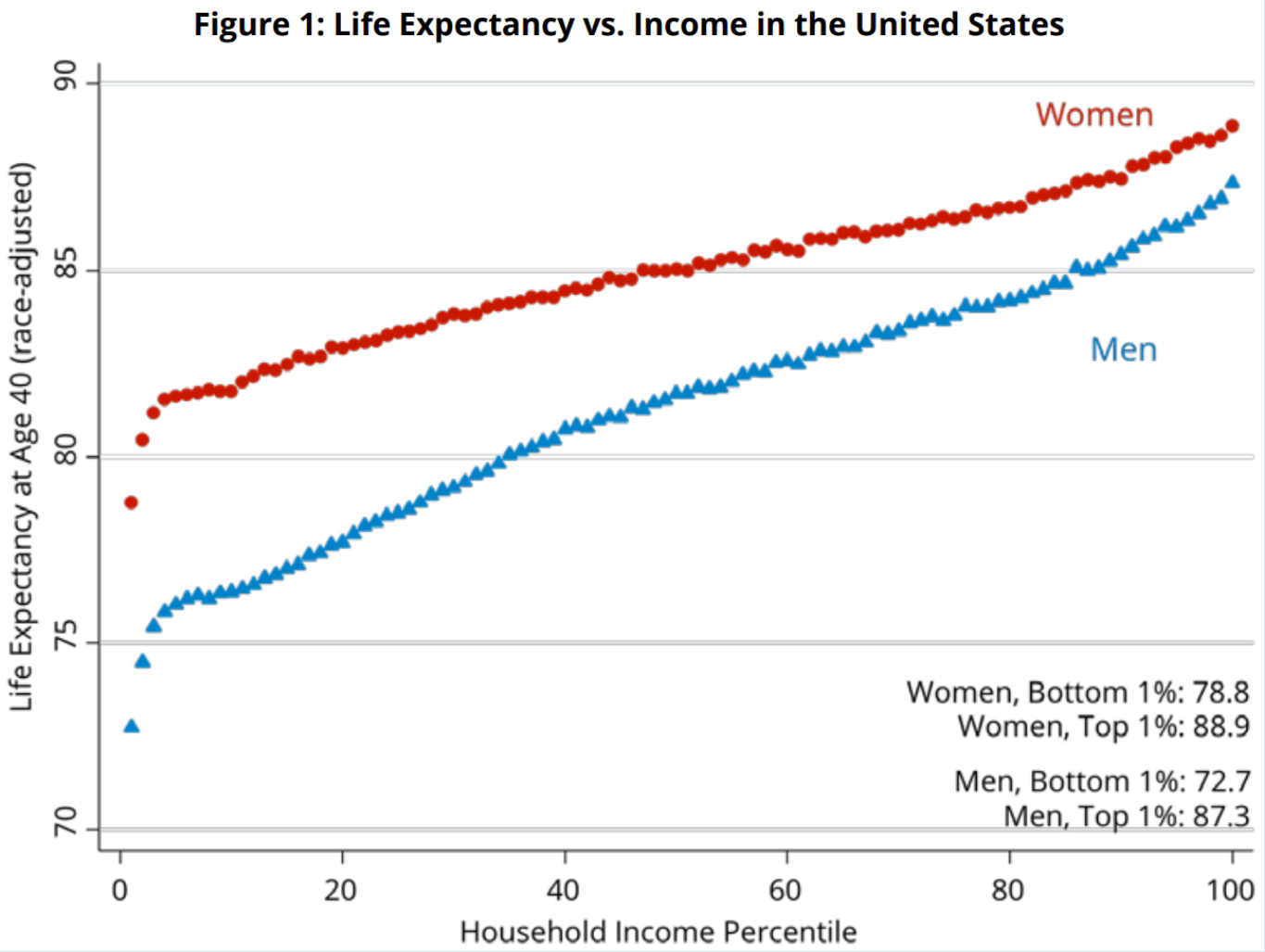
Looking beyond averages

Low-income men in the US have ≈ LE as Zimbabwean men!





Is this a causal relationship?

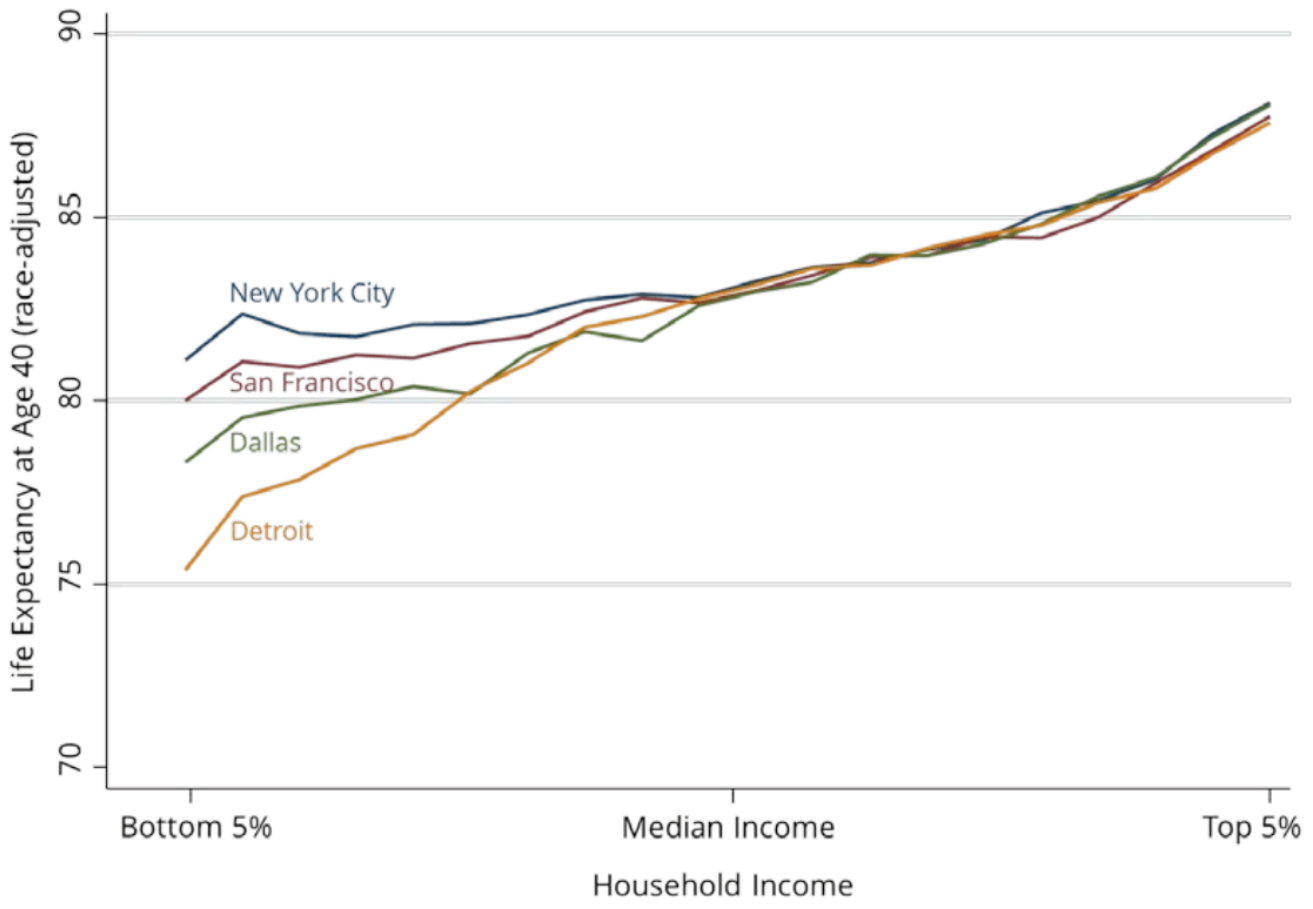


Does this suggest that giving more money to those at the bottom of the income distribution would ↑ their life expectancy?

Source: Chetty et al. (2016)

Descriptives can be incredibly powerful!

Place matters!



It's better to be poor in New York than in Detroit

Source: Chetty et al. (2016)

Discrete or categorical?

A practical rule

Discrete: takes a finite set of values

Continuous: takes a continuum

But every column in every dataset has finitely many values

- PISA `escs`, the socio-economic index: 62 distinct values in 12,136 rows
- PISA `math` score: 3,858 distinct values

Ask what the variable measures in the world, not how many values the column has

Three kinds of variables in the world

In the world	Example	Arithmetic makes sense?
A count	number of siblings: 0, 1, 2, ...	yes: an average of 1.4 siblings
A measurement	math score, socio-economic index, height	yes: any value in a range, rounded when recorded
A label	degree enrolled in, suburb, gender	no: there is no average suburb

Watch out for numeric-looking codes

Education recorded as

- 1 = completed primary
- 2 = completed high school
- 3 = completed a degree...

`df['education'].mean()` returns 2.3

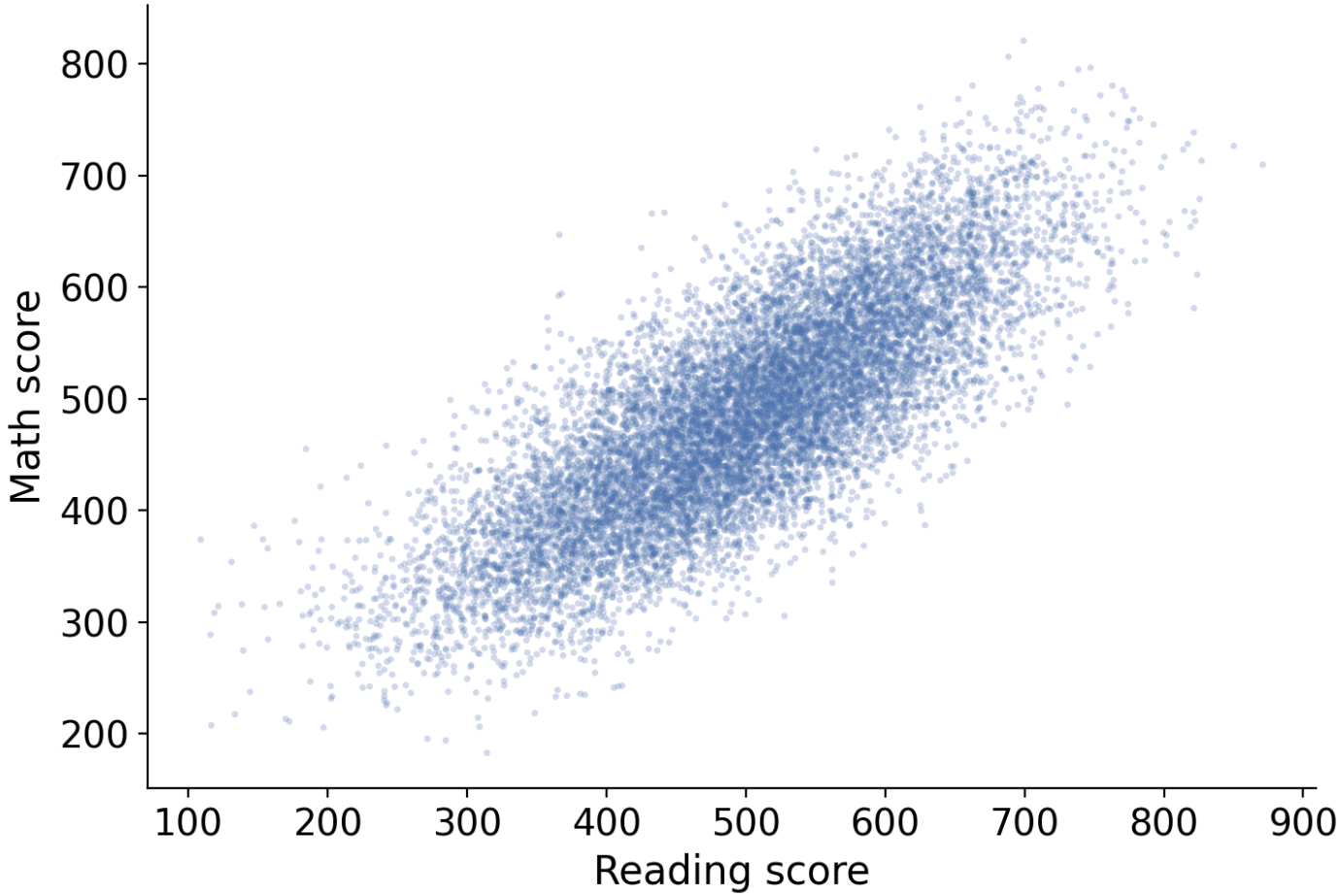
2.3 is a fact about the codes, not about anyone's education

Correlations

Six correlations

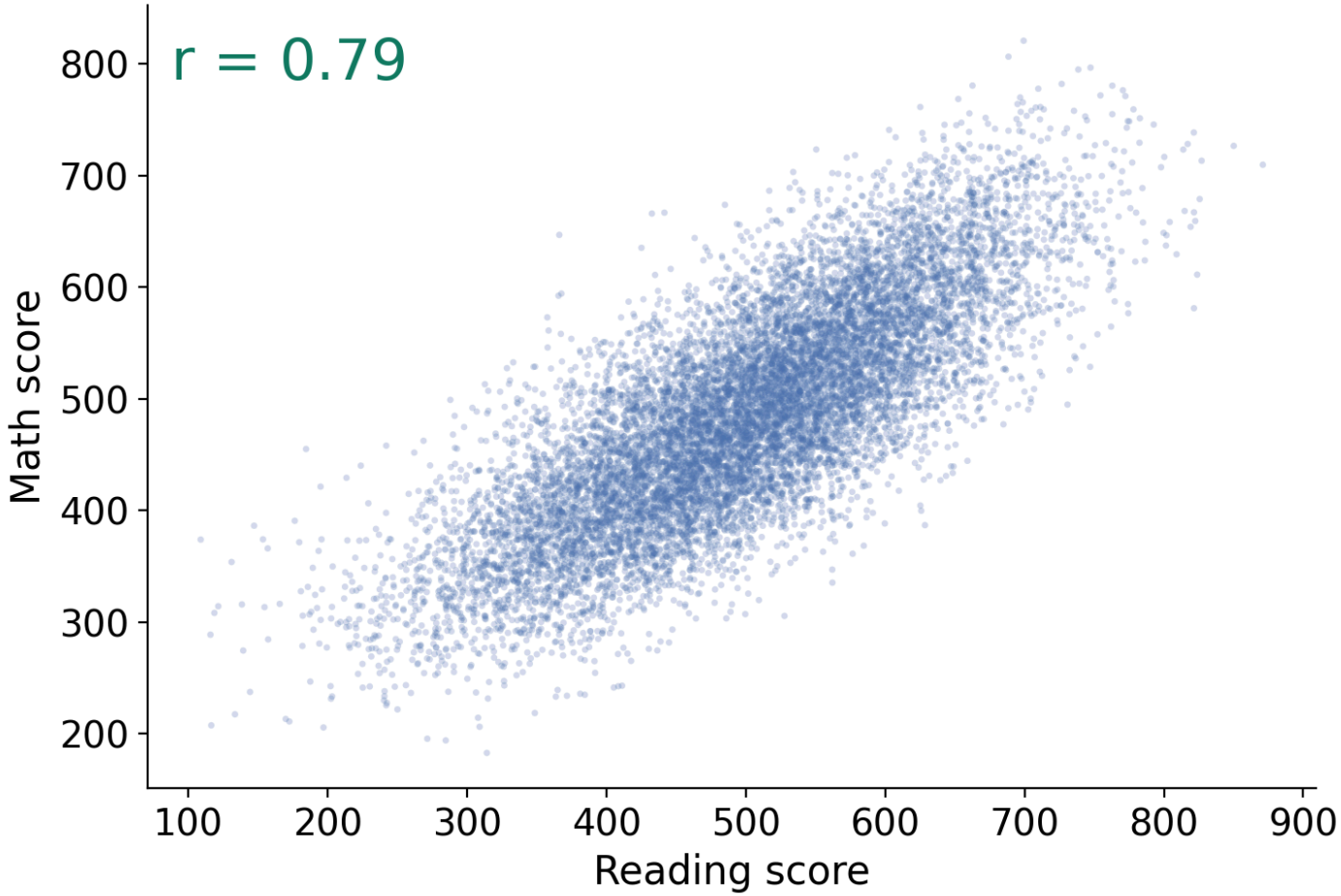
For each one: guess the correlation

Gallery 1: math & reading



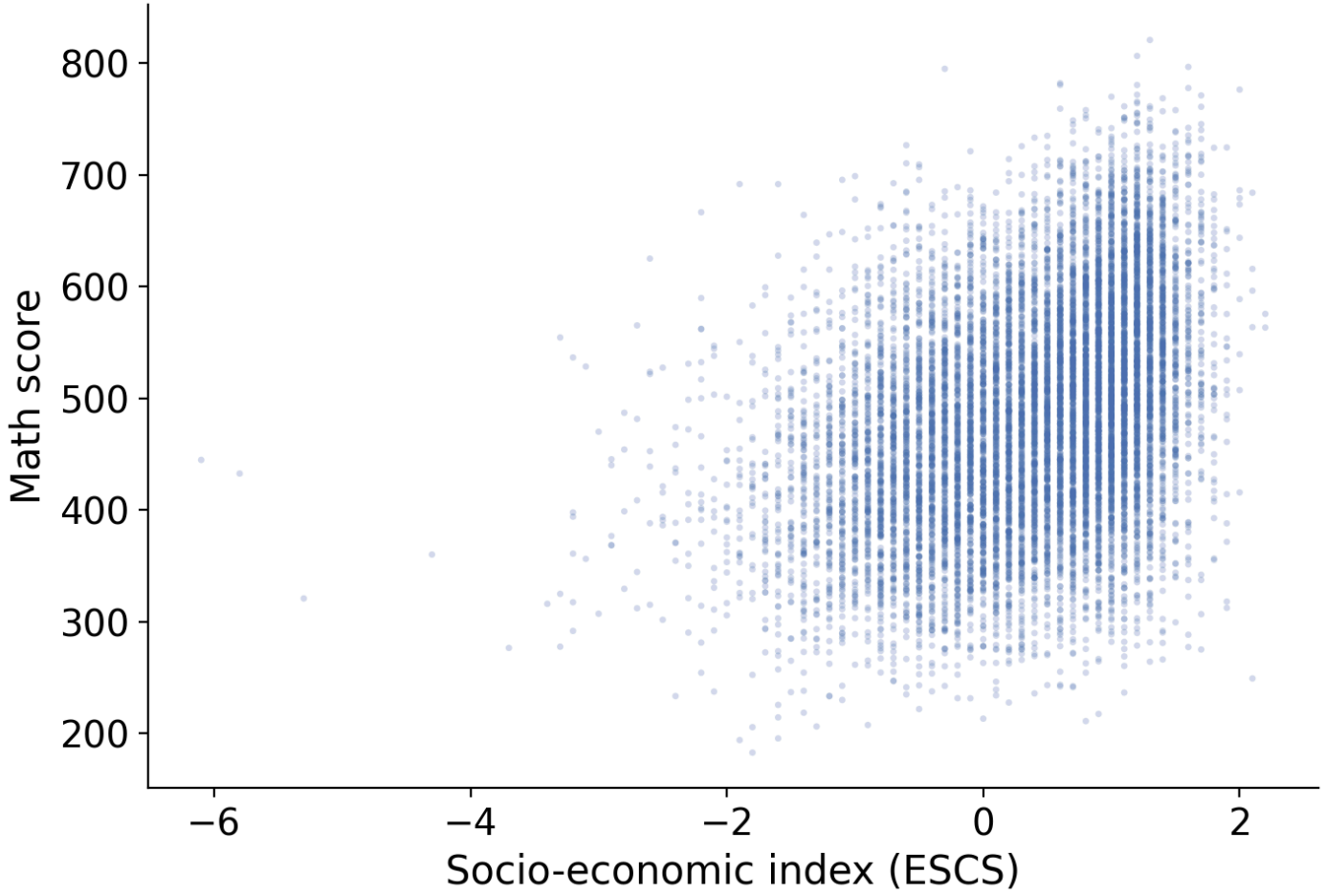
Data: PISA 2022, Australia (OECD)

Gallery 1: math & reading



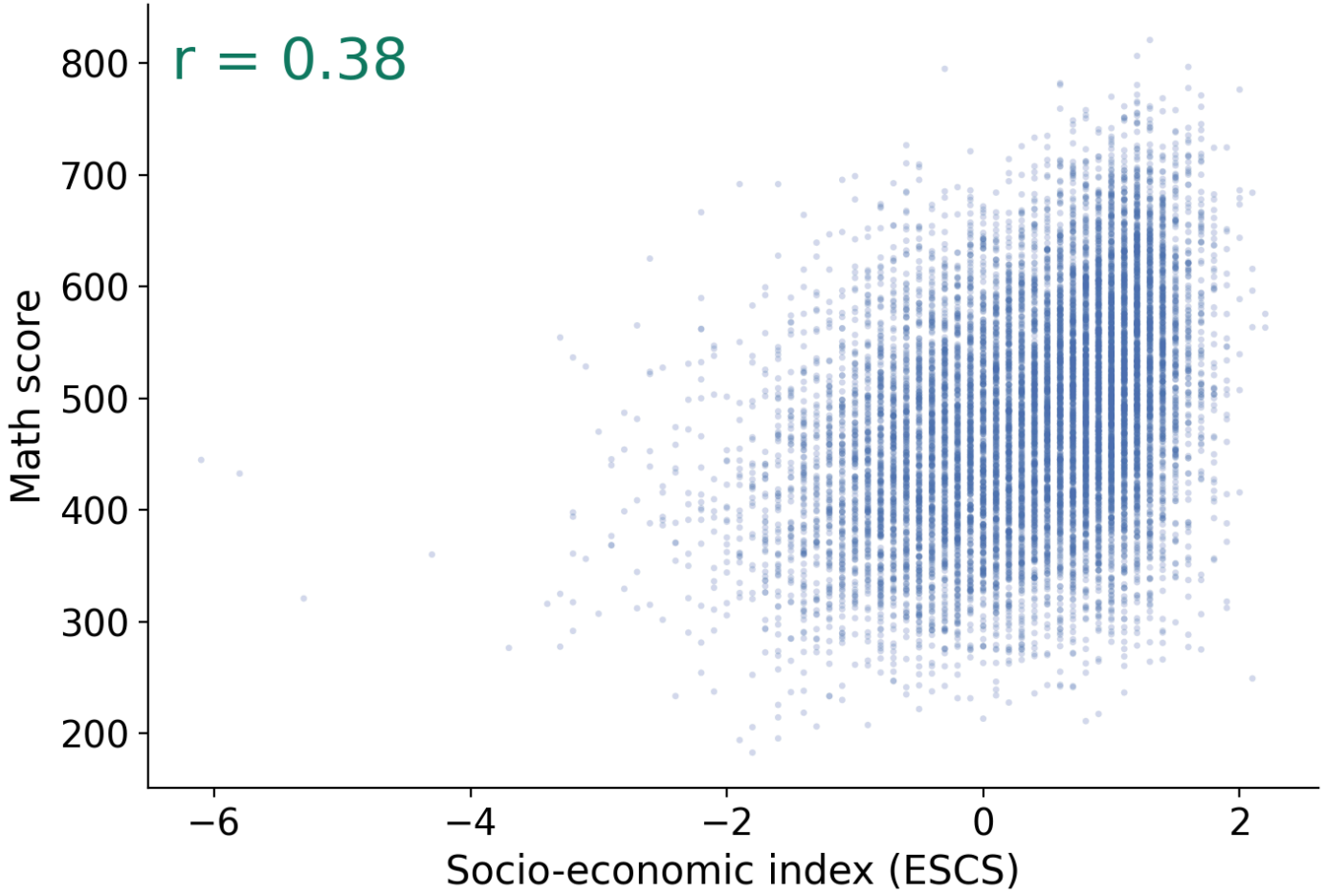
Data: PISA 2022, Australia (OECD)

Gallery 2: math & family background



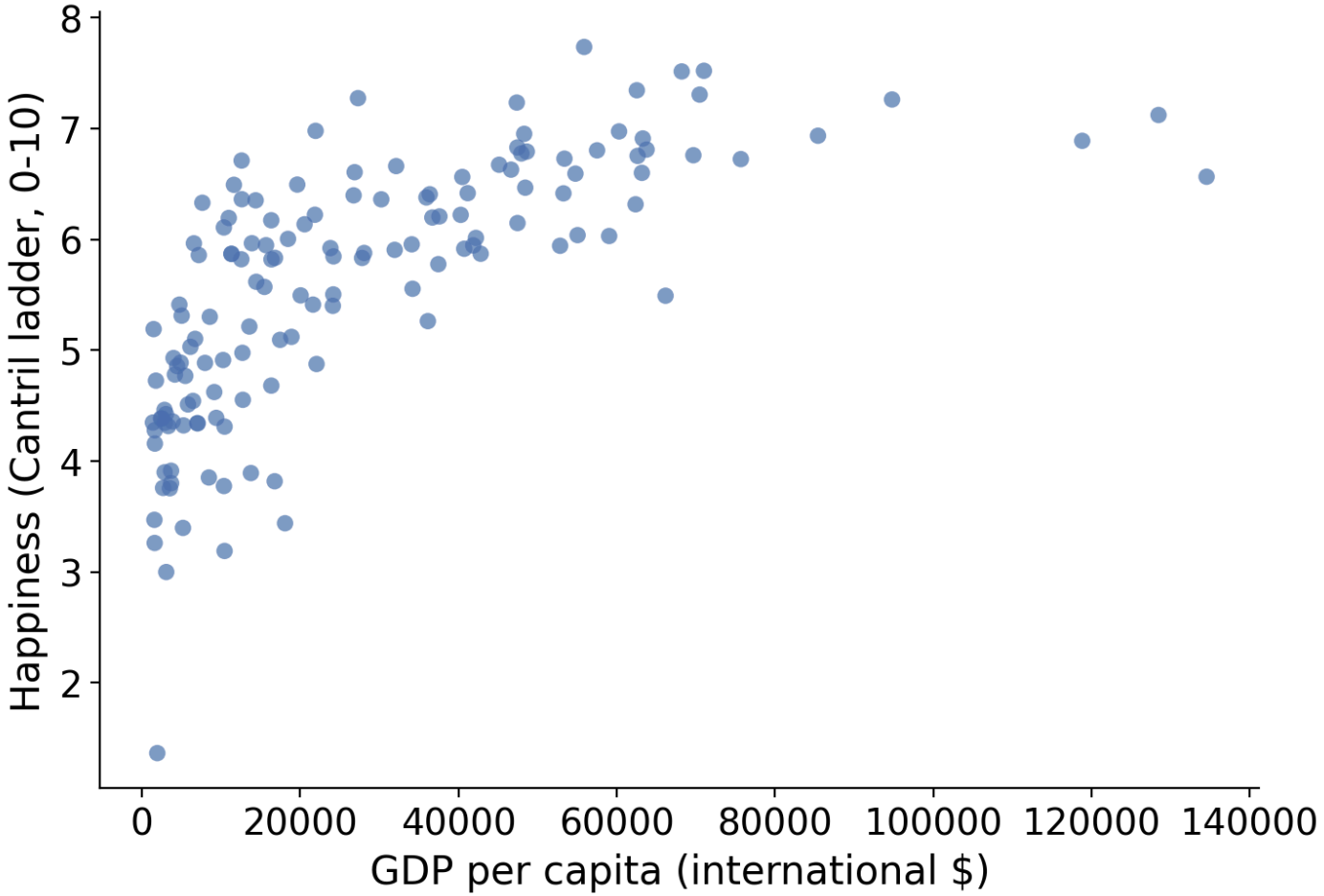
Data: PISA 2022, Australia (OECD)

Gallery 2: math & family background, 0.38



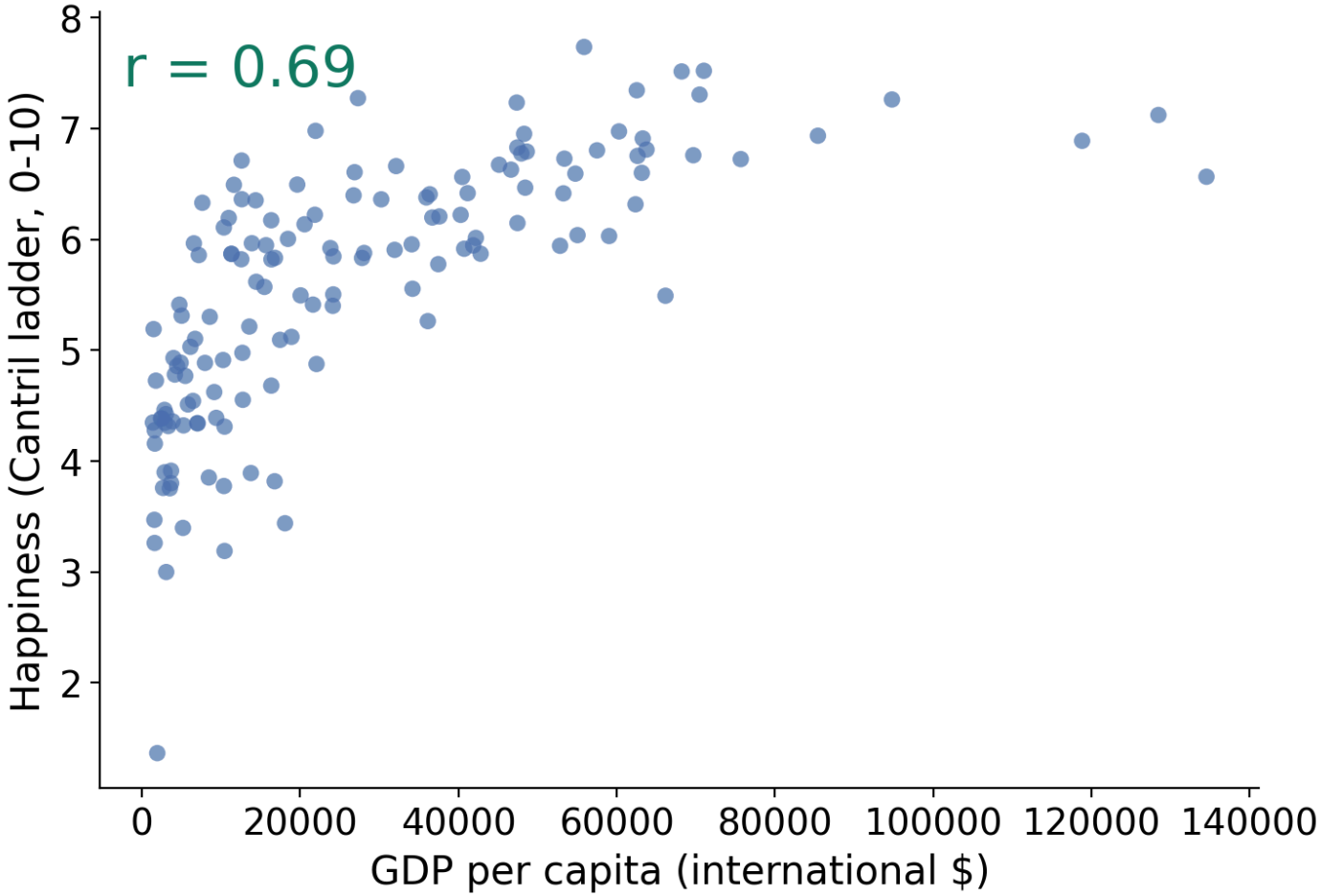
Data: PISA 2022, Australia (OECD)

Gallery 3: happiness & GDP



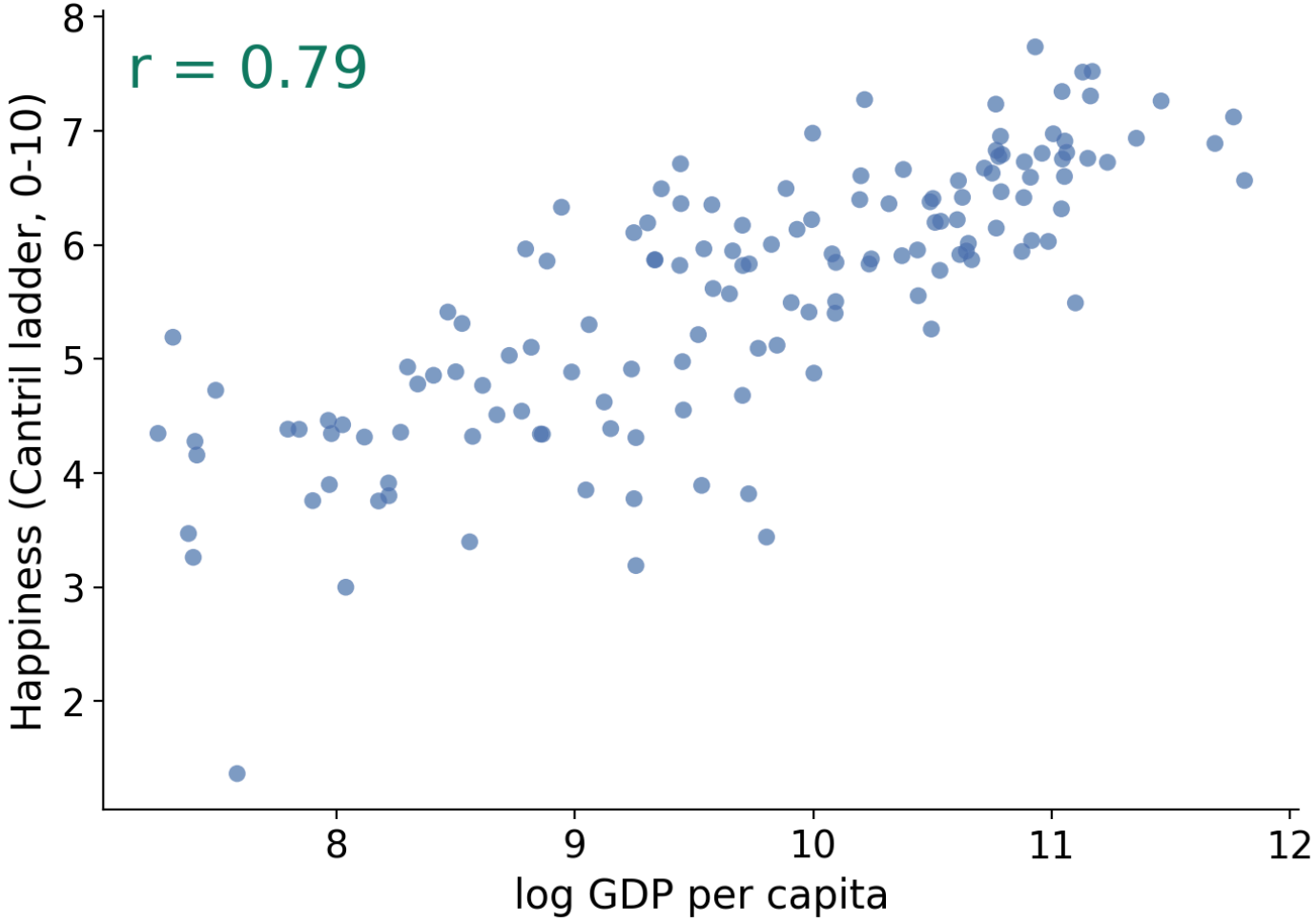
Data: Gallup World Poll & World Bank, via Our World in Data

Gallery 3:happiness & GDP



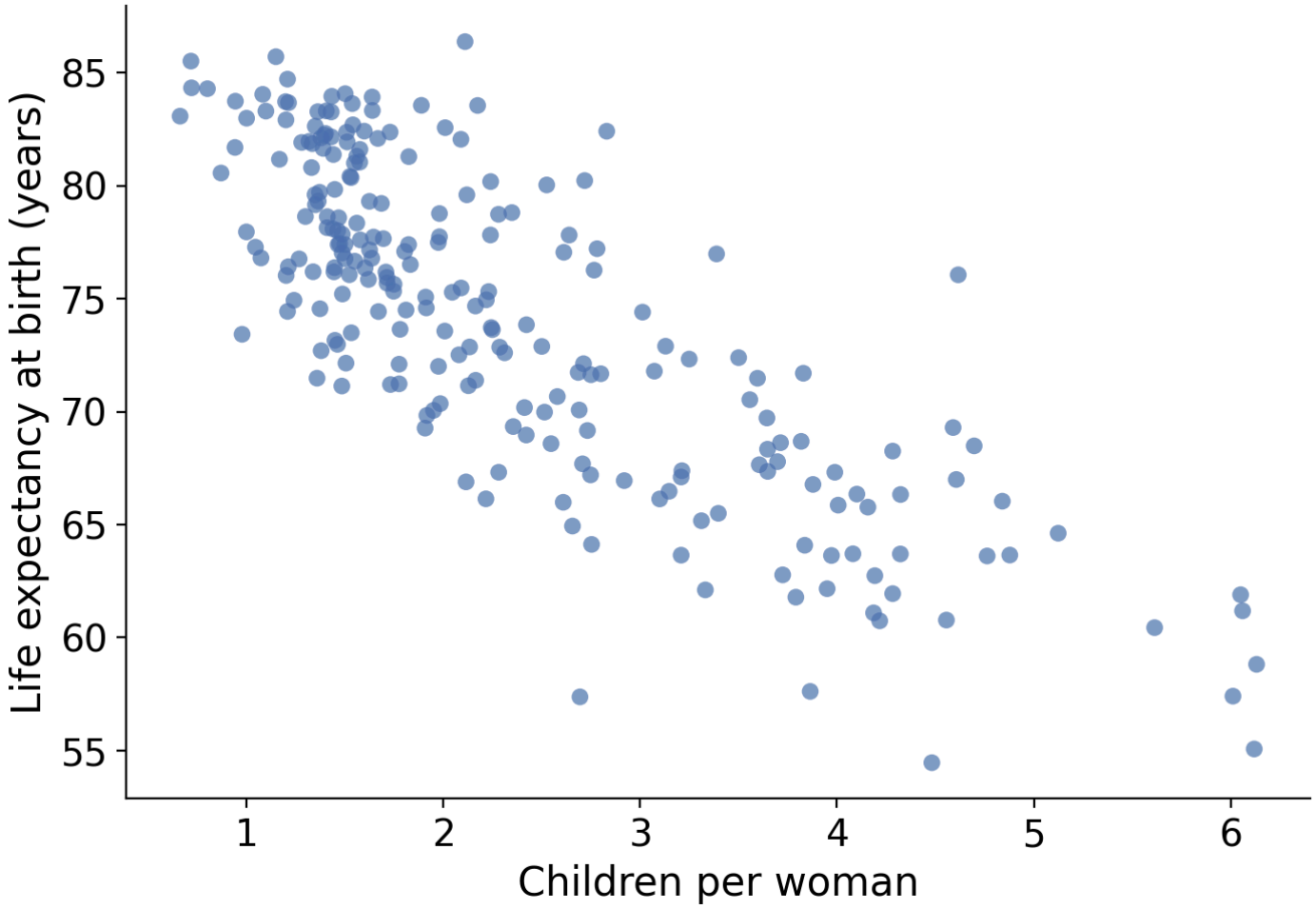
Data: Gallup World Poll & World Bank, via Our World in Data

Same countries, log GDP: r = 0.79



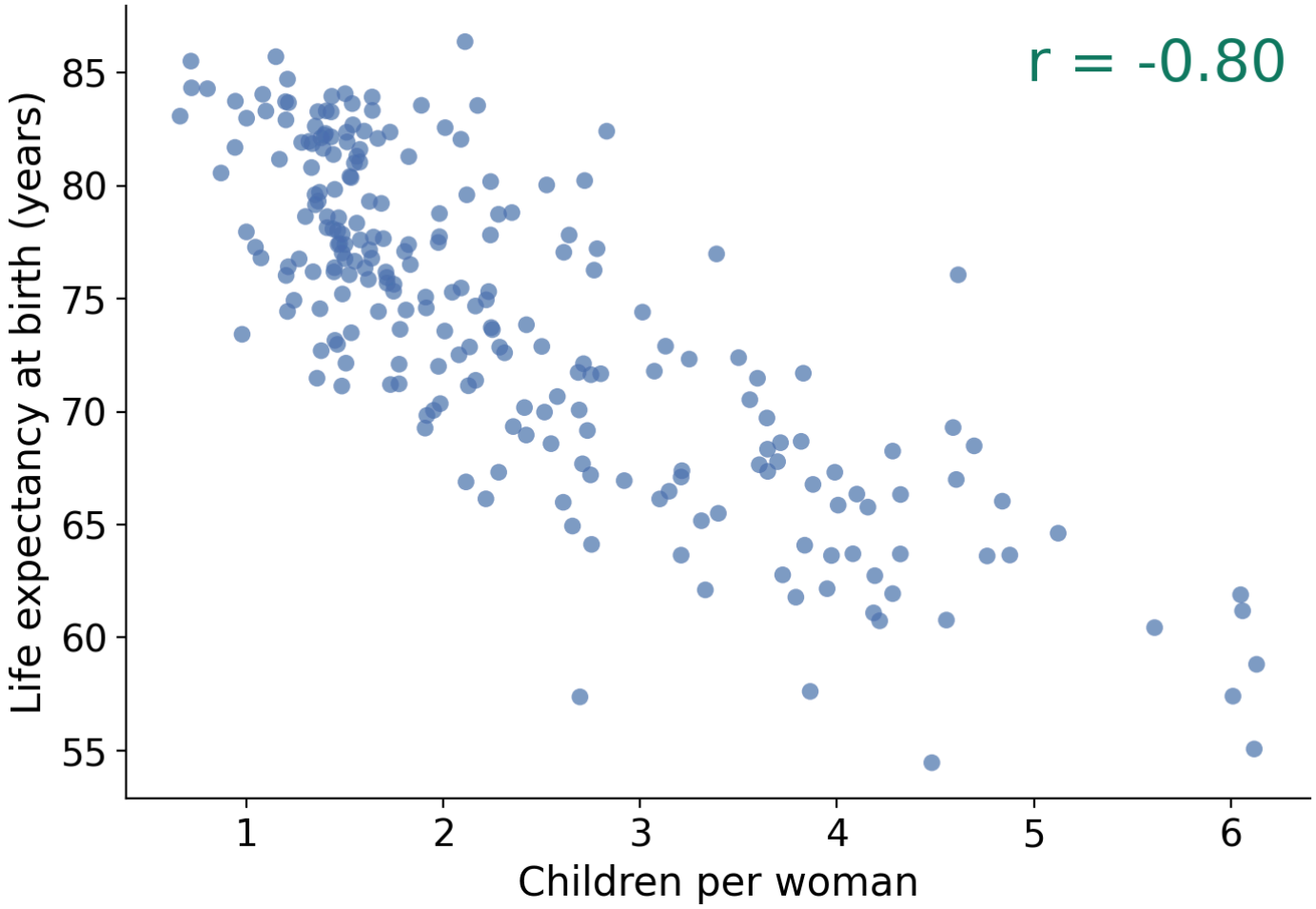
Data: Gallup World Poll & World Bank, via Our World in Data

Gallery 4: fertility & life expectancy



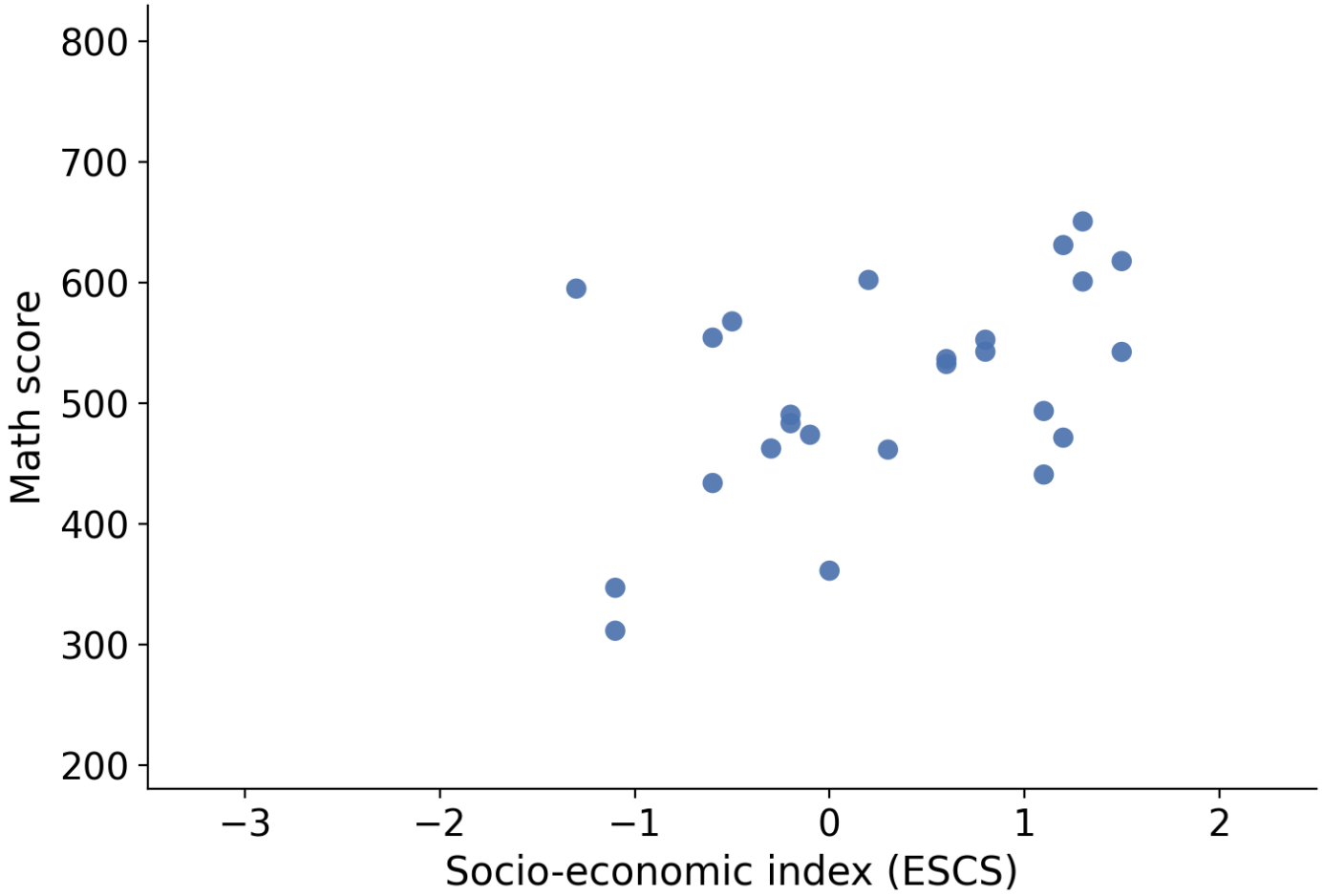
Data: UN World Population Prospects, via Our World in Data

Gallery 4: fertility & life expectancy, −0.80



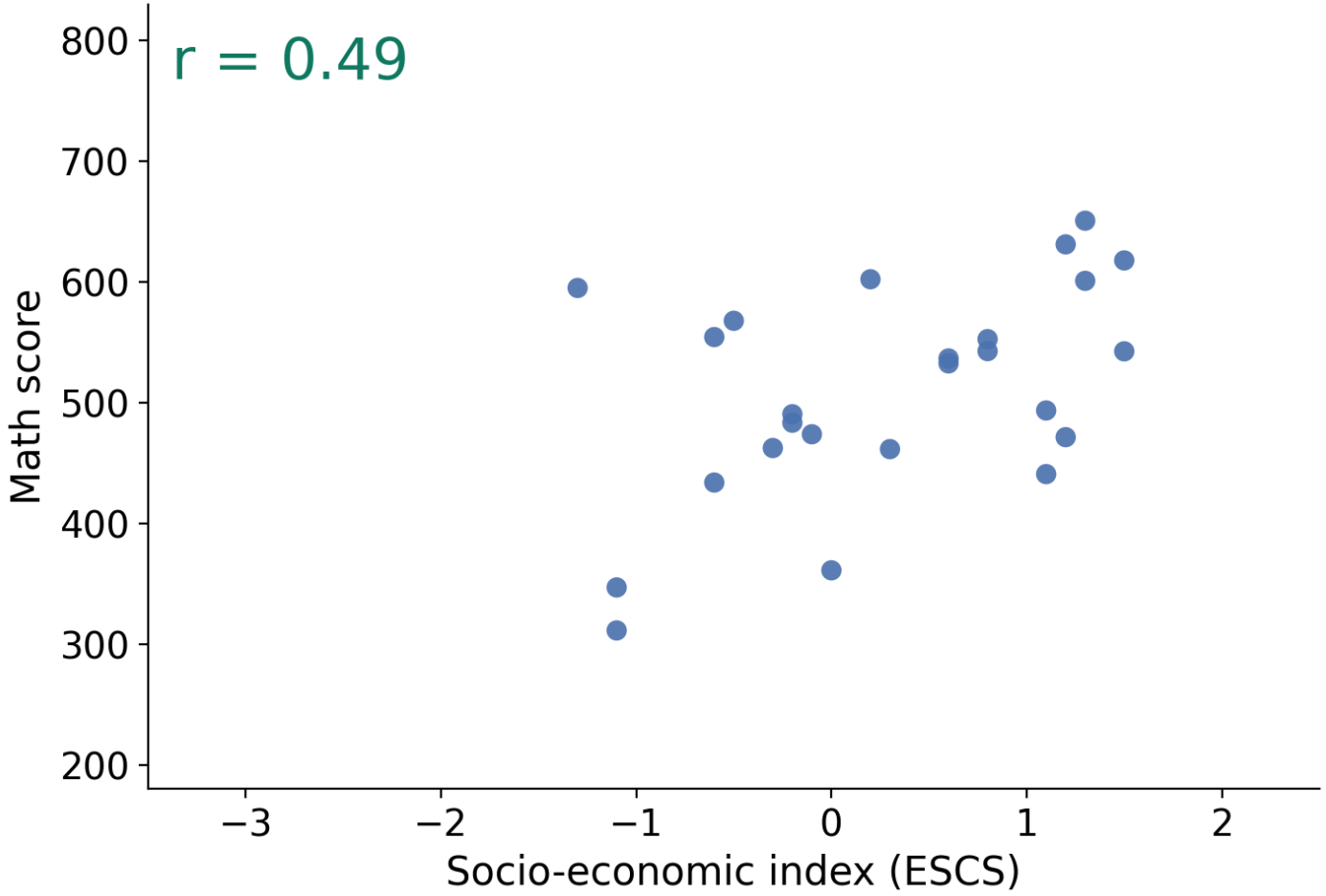
Data: UN World Population Prospects, via Our World in Data

Gallery 5: math & family background, $N = 25$



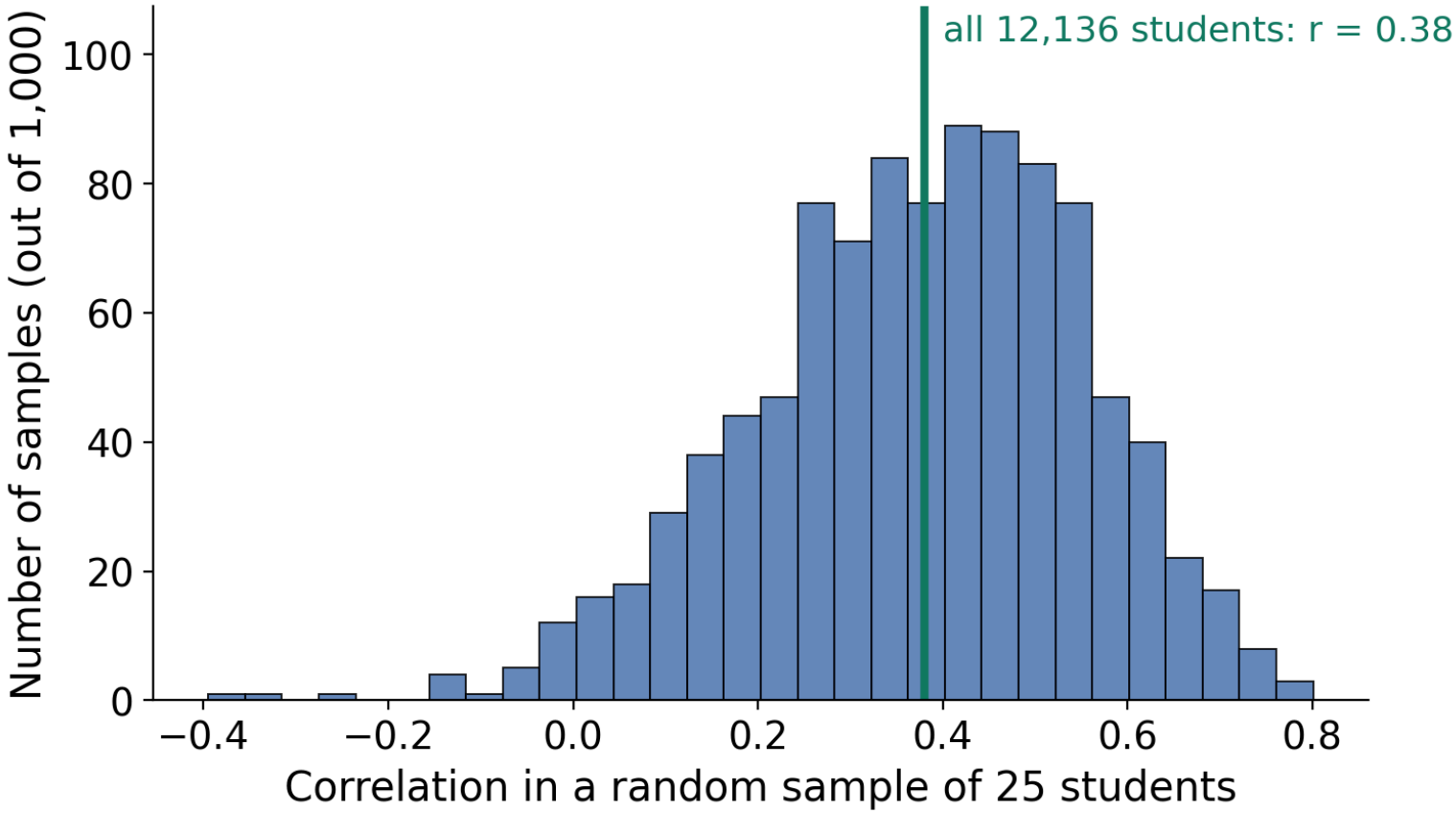
Data: PISA 2022, Australia (OECD), 25 students drawn at random

Gallery 5: 25 students, r = 0.49



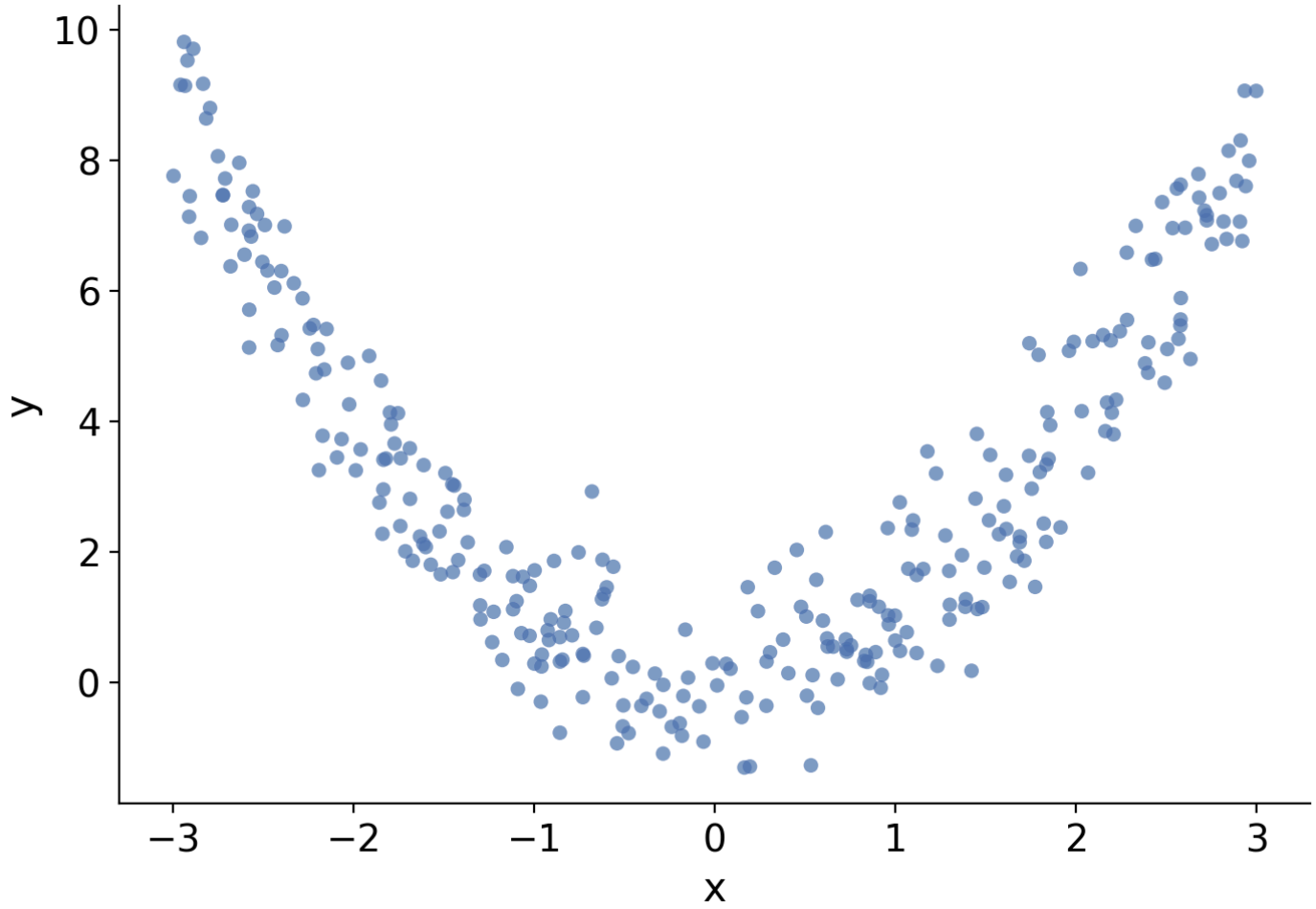
Data: PISA 2022, Australia (OECD), 25 students drawn at random

A thousand samples of 25: −0.4 to 0.8

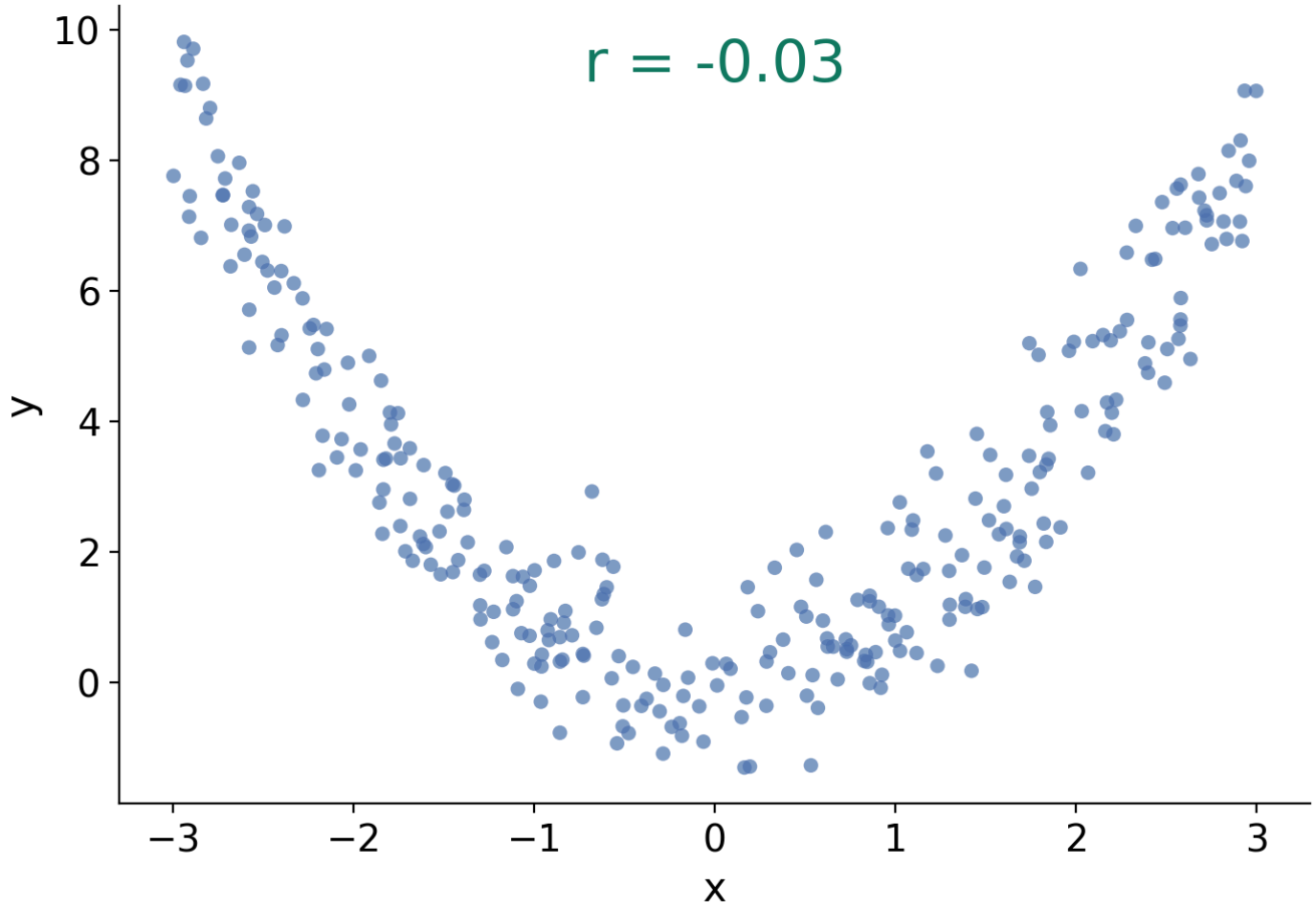


Data: PISA 2022, Australia (OECD)

Gallery 6: one more

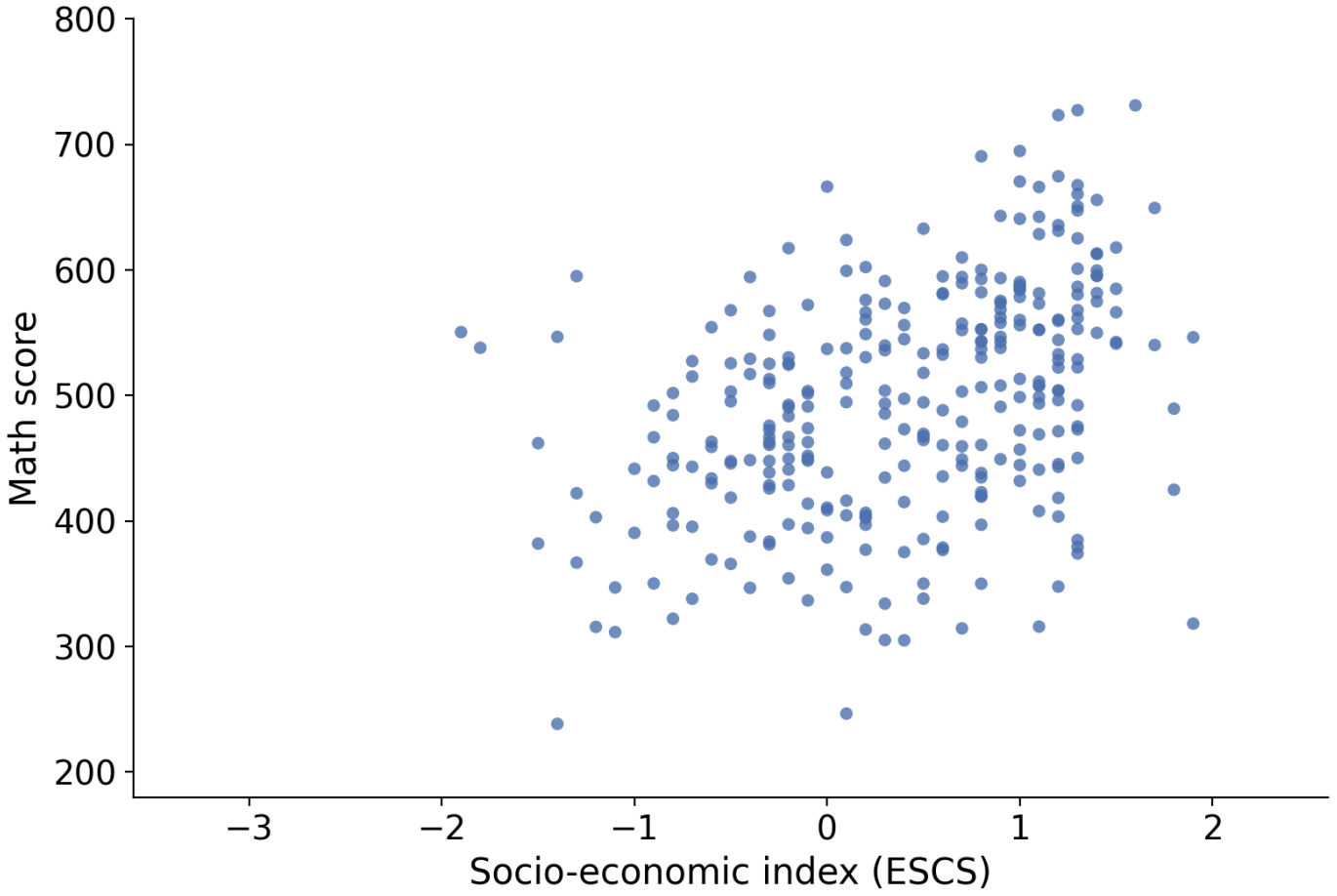


Gallery 6: r = −0.03, yet y depends on x



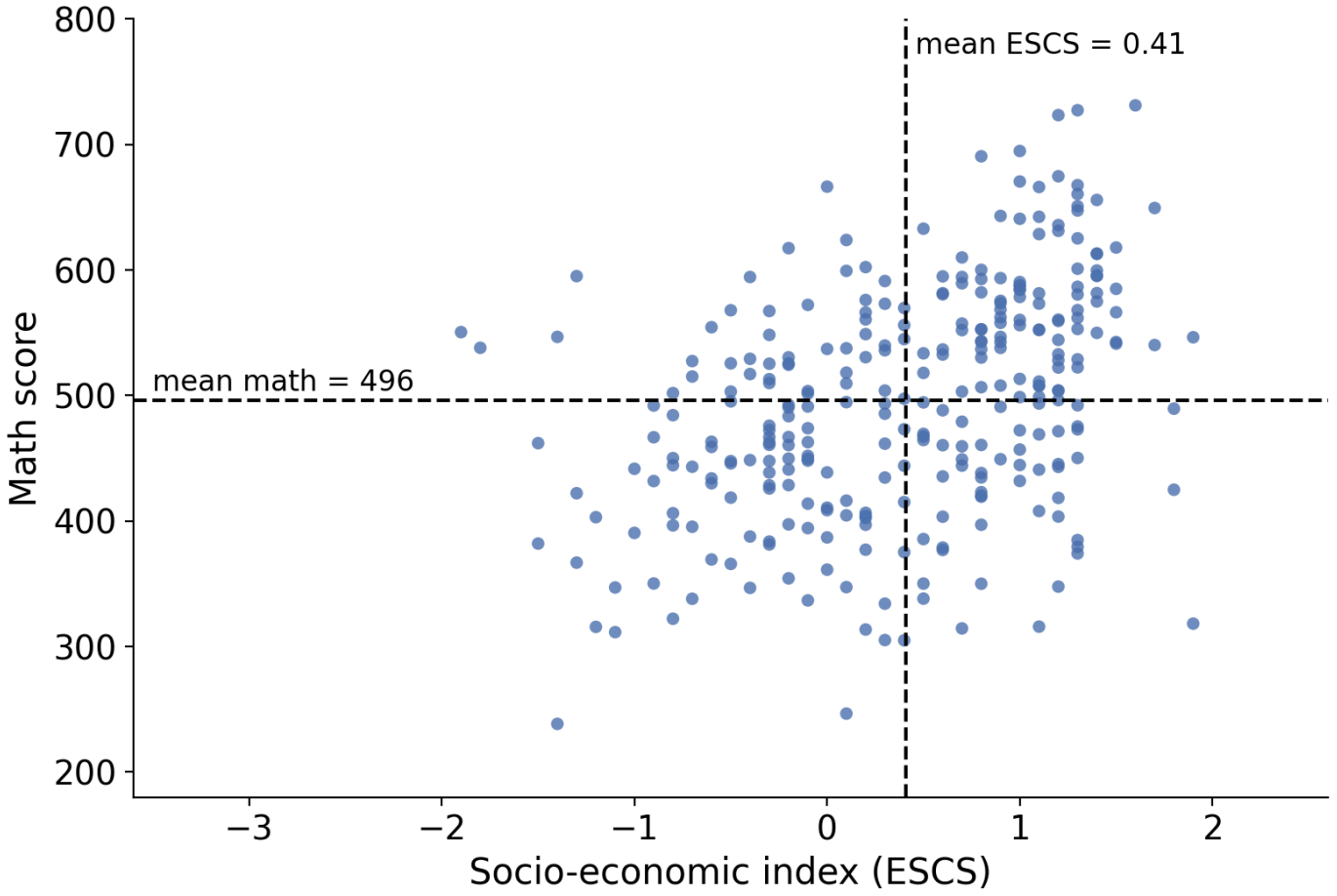
Where a covariance comes from

300 students, math against family background



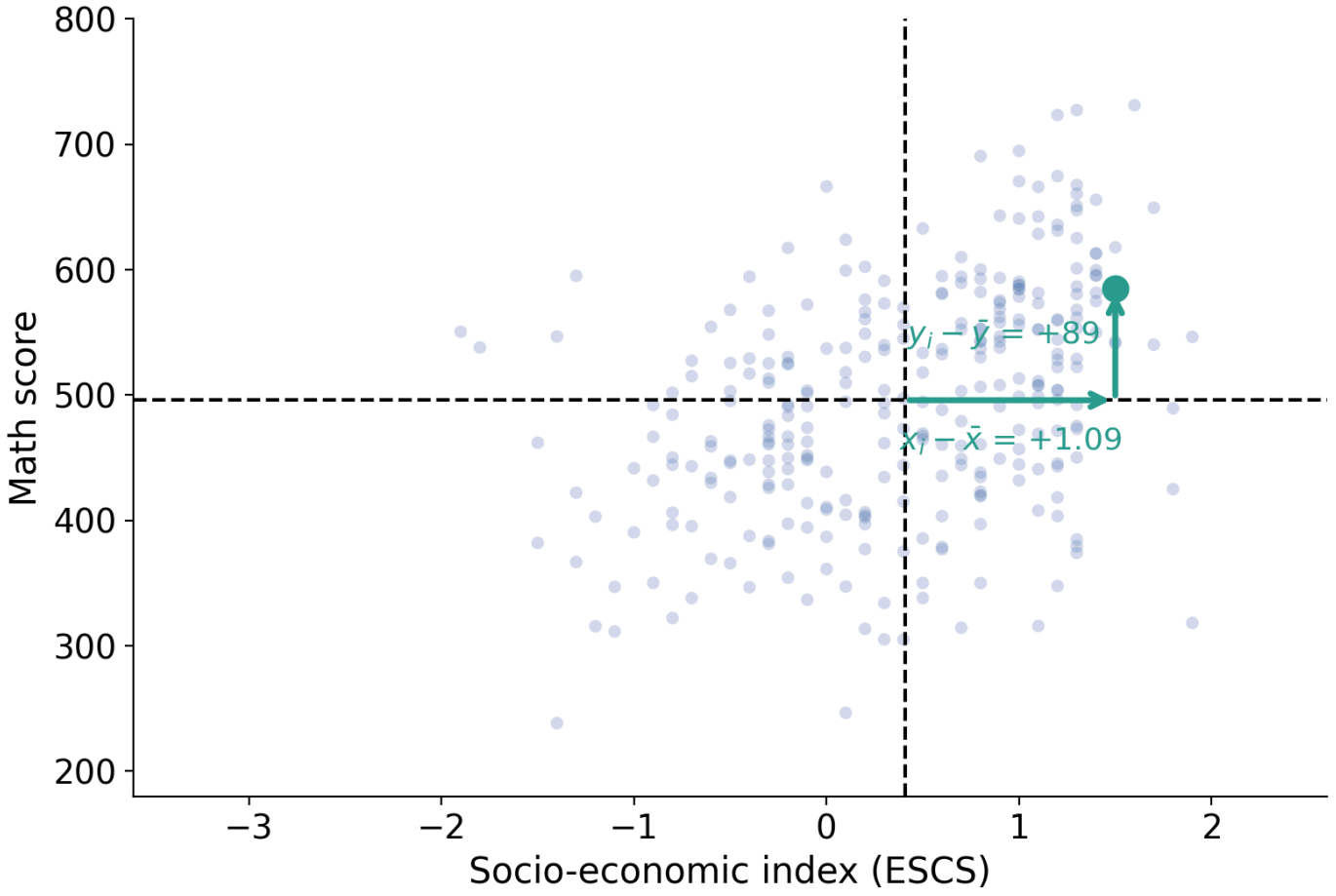
Data: PISA 2022, Australia (OECD), 300 students drawn at random

Step 1: find the two means

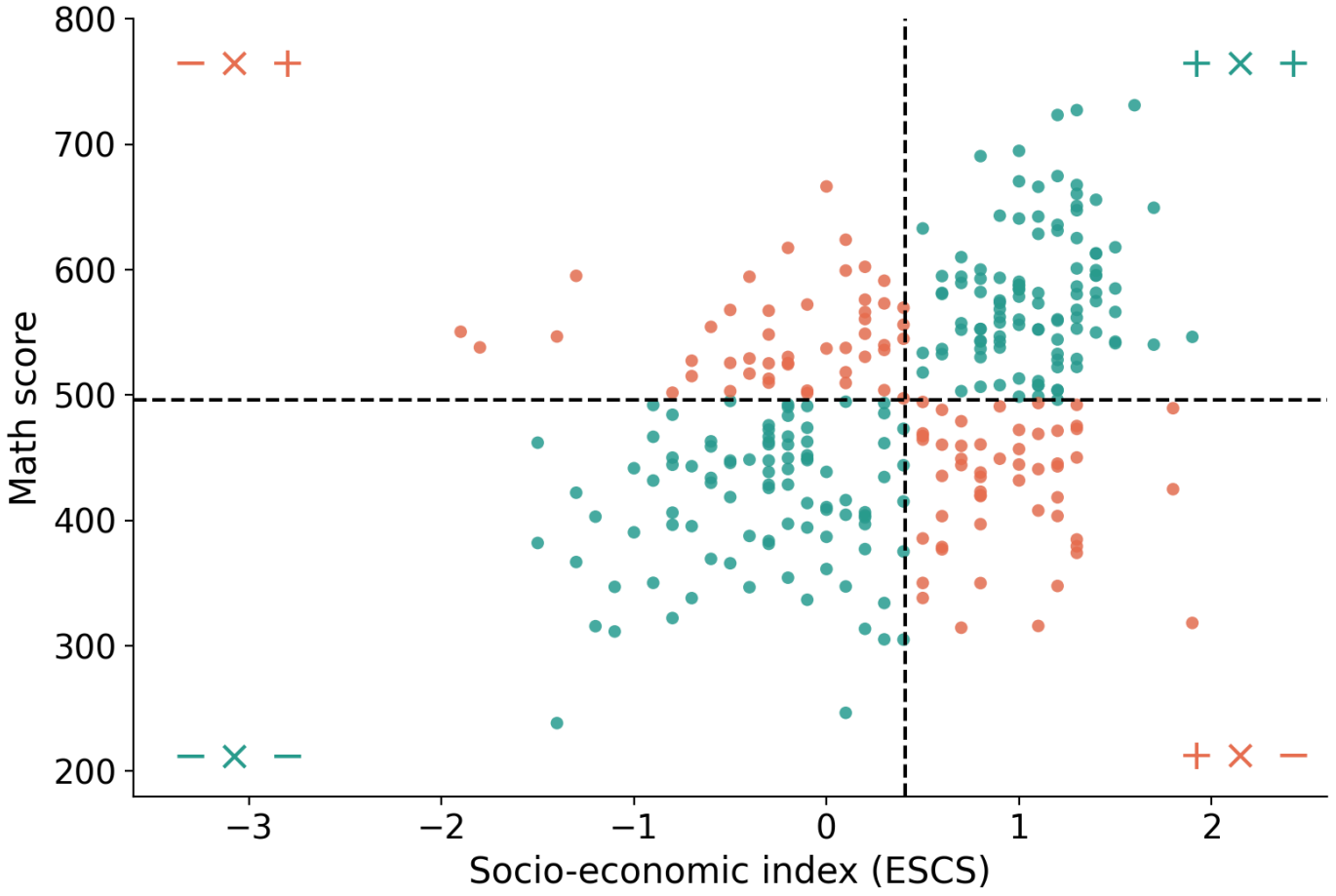


Data: PISA 2022, Australia (OECD), 300 students drawn at random

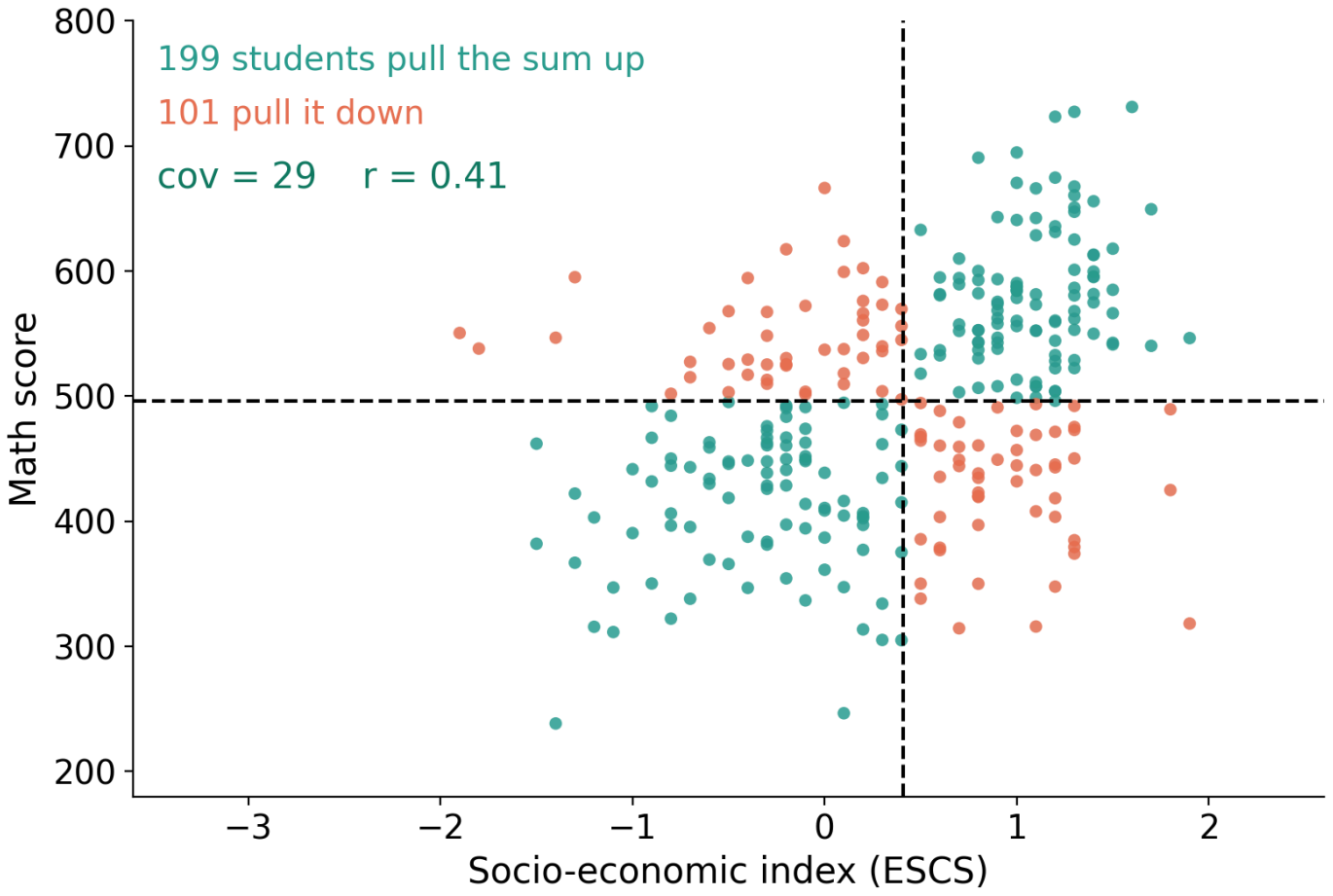
Step 2: each student has two deviations



Step 3: the sign of the product is the quadrant



Step 4: add them up



Nonsense units, so we standardize

$$\operatorname{cov}(X, Y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 29 \text{ index points} \times \text{score points}$$

$$r = \frac{\operatorname{cov}(X, Y)}{s_X \, s_Y} = \frac{29}{0.78 \times 91.7} = 0.41$$

Some take-aways

- A conditional expectation is a function: one average for every value of X
- Averaging hides the ends: 14.6 years, inside one country
- Correlation sees straight lines only, and a sample's r is itself random
- The covariance is a vote by quadrant; dividing by the SDs removes the units