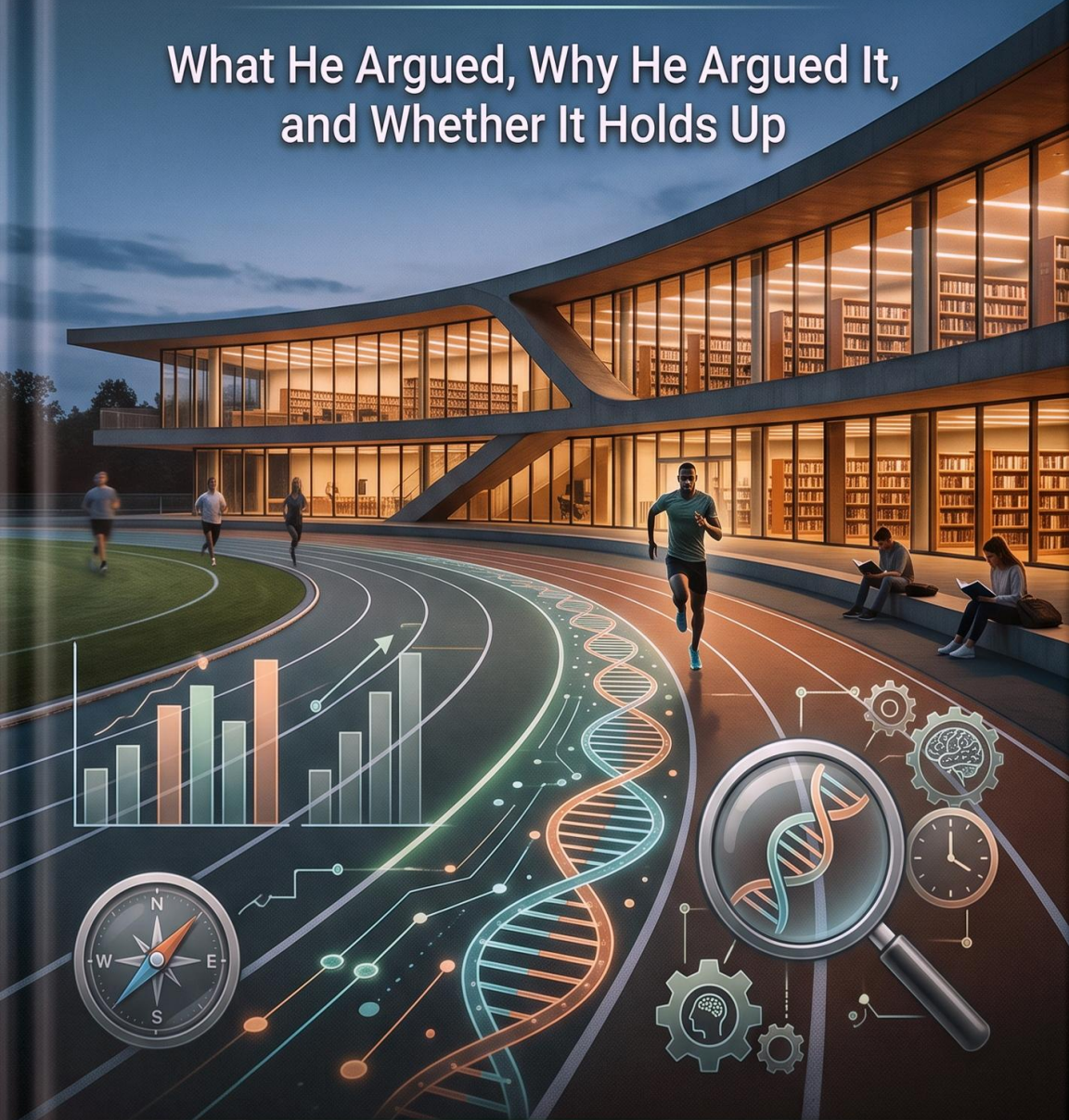


UNOFFICIAL GUIDE TO THE IDEAS OF DAVID EPSTEIN

What He Argued, Why He Argued It,
and Whether It Holds Up



Written by Claude Code | Instructed by V.Korostyshevskiy

Unofficial Guide to the Ideas of David Epstein

What He Argued, Why He Argued
It, and Whether It Holds Up

September 2nd 2026

Written by Claude Code | Instructed by V.Korostyshevskiy

TABLE OF CONTENTS

Preface.....	2
Is There Such a Thing as a Sports Gene?	4
Does the Same Training Work on Everyone?	9
Can You See It Coming?.....	15
Why Do the Bodies Keep Getting Stranger?.....	20
Does Ancestry Explain Who Wins?	25
Who Counts as a Woman?	32
What Is Genetic Testing Actually For?.....	39
Did the Ten-Thousand-Hours Argument Ever Get Settled?	45
When Does Experience Actually Make You Better?.....	49
What Makes Learning Stick?	55
Is Grit a Thing?.....	60
Can Anyone Forecast Anything?.....	65
Should You Keep Your Options Open?	71
Does Breadth Actually Pay?.....	76
Do Constraints Make You Better?	81
What Should You Settle For?.....	87
So What Is He, Then?.....	92
Endnotes.....	97

PREFACE

This book has one reader.

It was not written for a market, and there is no second copy waiting for a general audience. It exists because you wanted to know what David Epstein has been arguing across three books, why he was arguing it, and whether any of it survives contact with what the relevant fields hold now.

So a few things about how it was made, because you should know what you are holding.

The narrator is software. I have not run, jumped, trained, or had a body. What I have done is read all three books end to end, pull out every load-bearing claim I could find, and check each one against the current literature. Where this book says *I*, that is what the word refers to. There is no biography behind it and I am not going to invent one.

Anything he retracted himself is not in here. Where a later book of his abandons or narrows a position from an earlier one, the abandoned version does not appear at all, not even as a correction. He did that work already, in public, and auditing a man on a position he withdrew is scoring points rather than checking facts. The biggest single exclusion is the afterword to the paperback of *Range*, where he concedes a good deal to his critics. None of that is audited here.

The benchmark is what the relevant field holds now, and nothing else. Not whether his three books agree with each other, and not whether his 2013 footnotes supported his 2013 claims. "His evidence was thin" is not a finding in this book. "The field now holds otherwise" is.

Every chapter does the same four things: what he wrote, why it made sense when he wrote it, whether it holds up, and the story he told to make it stick. Nothing else.

Verdicts come in five kinds. A claim can be **vindicated**, meaning the field moved toward him. **Mainstream**, meaning it was already the consensus, or has quietly become it. **Contested**, meaning the field is genuinely split, in which case I describe the split rather than resolving it. **Superseded**, meaning the field moved away or

reframed the question. Or **untested**, meaning nobody has checked, which is a real answer and appears more often than any of the others.

On numbers. There are around thirty figures in his books I could not independently source. Where one appears here, I say so in the sentence rather than hiding it in a note, and I distinguish three cases: he named no source, he named a source that has never been published, or he named a real source whose full text I could not obtain. The first two are facts about his books. The third is a limit on this one.

On the sourcing behind this. The audit ran to about 167,000 words of research across twelve fields, with 668 citations, every one of them resolved against a public registry before it was trusted. The chapters carry only a handful of notes each, attached to things that surprised me or that you might want to check. Everything else is in the working files.

And a warning about what I expected to find. I expected to catch him. That is what an audit is for, and the appetite is real even when the intention is honest.

Of three hundred and twenty verdicts, thirteen have been overtaken.

He is almost never wrong. He is frequently unchecked, which is not the same thing and turns out to be considerably more interesting. The last chapter is about the difference.

IS THERE SUCH A THING AS A SPORTS GENE?

At the 1964 Winter Olympics a Finnish cross-country skier won the fifteen-kilometre race by forty seconds.

That number means nothing until you hear the other one. The five men who finished behind him were all within twenty seconds of each other. He had beaten the field by twice the margin that covered the entire rest of the podium, in an event where nobody has managed it before or since. Two days later he won the thirty-kilometre race by over a minute.

For most of the next thirty years, people assumed Eero Mäntyranta had cheated. His blood carried up to sixty-five percent more red blood cells than an average man's. Red blood cells carry oxygen, oxygen is the binding constraint in endurance sport, and injecting something to raise your red cell count is the oldest form of blood doping there is. He looked exactly like a man who was doing that.

Then, two decades after he retired, Finnish haematologists went to look at his family.

They started with his bone marrow. The standard test is to add the hormone that tells marrow to produce red blood cells, and then watch it respond. They never got that far. Mäntyranta's marrow began producing red blood cells before anybody stimulated it with anything.

The gene that controls this is a receptor, and the receptor has a built-in brake: a segment at its tail end whose job is to switch the signal off once enough cells have been made. That gene is about seven thousand letters long. In Mäntyranta, one letter was wrong. The change inserts a premature stop, so the receptor gets built without its final fifteen percent, which is the part with the brake on it.

The signal to stop never arrives properly. Production simply keeps running.

They examined ninety-seven members of the family. Twenty-nine of them carried it, and it did not appear to harm any of them. When the geneticist sat with three elderly Mäntyranta women, two with the mutation and one without, it was the woman without it who had all the health problems.

Two of the carriers were his sister's children. One won a world junior title. The other won an Olympic gold. Nobody else in the family raced.

Mäntyranta himself denied to the end that any of this had helped him. He said it was determination. The geneticist who found the mutation disagreed flatly, and noted that the man's haemoglobin was the highest he had ever seen in a human being.

And there is a detail that has to sit next to all of that, or the story means the wrong thing. As a child, Mäntyranta skied to school across a frozen lake, an hour each way, in the dark, because that was the route. He once went through the ice. Skiing was not his sport. It was his commute.

So this is a man with a mutation that manufactures the exact physiological advantage his event rewards, who also happened to spend his entire childhood doing enormous volumes of the specific activity. Both things are true at once, and the book's own conclusion at the end is that he is one hundred percent nature and one hundred percent nurture, which sounds like a dodge and is actually the only defensible reading.

What the book is arguing

That is the closing chapter of *The Sports Gene*, and it is there as an existence proof. One gene. Large effect. Documented through a family tree. Real athletic consequences.

But a proof of existence is not a proof of frequency, and the book knows it. Mäntyranta is presented as extraordinarily rare. The rest of the book goes looking for the ordinary version: common genetic variants, carried by millions of people, each nudging performance a little.

The flagship candidate was a gene called ACTN3, which builds a protein used in fast-twitch muscle fibre. It comes in a working version and a broken one, and the broken one is common: roughly a fifth of Europeans carry two copies of it and make none of the protein at all. Almost no elite sprinter is in that group. In one early study, not one of thirty-two Olympic sprinters was.

The chapter reporting this is careful in a way that turns out to matter enormously, and I will come back to it.

Why it made sense to say at the time

The human genome was sequenced in 2003, and the expectation was that the answers would fall out of it within a few years. They did not. Reading three billion letters turned out to be much easier than working out which of them mattered.

So the field did the sensible-looking thing. Pick a gene you have a solid biological reason to suspect. Compare a few hundred people who have the trait against a few hundred who do not. See whether the versions differ. If they do, you have found something.

This is called candidate-gene research, and it was not a fringe method. For about fifteen years it was *the* method, across sports science, psychology and psychiatry alike. Thousands of papers.

The problem with it is a statistical one that almost nobody was taking seriously at the time. If the true effect of a single gene is very small, a study of a few hundred people cannot see it. That much was understood. What was not understood is the second half: such a study can still produce a result, because with a small sample the only way to clear the threshold for publication is for the random noise to happen to be pointing the right way. So an underpowered field does not produce faint, cautious findings. It produces confident ones that are mostly noise, and it publishes them.

The Sports Gene called part of this in 2013, while it was still happening. The book says the single-gene studies were failing because single genes have effects too small to detect in samples that size.

That is a good call, made early, and it is nearly right.

Does it hold up

Split down the middle, and the split is clean enough to state as a rule.

Every rare, large-effect claim in that book survived. Every common, small-effect claim shrank or died.

Mäntyranta is untouched. The mutation is real, the mechanism is understood, the family tree is documented, and nothing that happened to the rest of the field so much as brushed it. The same goes for the other large-effect example the book

uses, a mutation in a gene called myostatin, whose normal job is to limit muscle growth. Break it and the limiter comes off. It produces cattle so muscular they look computer-generated, whippets with double the normal muscle, and, in a handful of documented human cases, small children with visible abdominal definition and roughly twice the usual muscle mass. That work has held up completely.

The common-variant research did not hold up. And it failed for a reason worse than the one the book gave.

The book said the effects were too small to see in small samples. What the field actually found, once samples reached the tens and hundreds of thousands, was that the candidate genes as a *set* were no more associated with the traits they had been chosen for than genes picked at random.¹

That is a different and much more damaging result. Too-small-to-see implies the signal was there and the instrument was weak. What happened instead is that a generation of researchers had picked the wrong genes on plausible-sounding biological reasoning, then published noise for fifteen years, and the noise looked like signal because underpowered studies only ever publish their flukes.

ACTN3 is the interesting survivor. It came through well-powered pooling, which almost nothing from that era did. And what came through is that it explains under one percent of the difference in sprint times between elite sprinters.² The most recent meta-analysis of muscle function finds it associated with lifting, jumping and gripping, and not with sprinting at all.³

So the variant is real and the phrase attached to it is not. There is no gene for speed. There is a gene that does something, slightly, in the general vicinity of the thing.

Here is where the carefulness pays. *The Sports Gene* does not close that chapter in its own voice. It gives the last word to Kathryn North, the geneticist who discovered the variant, who says it looks like one gene contributing a little, that there may be hundreds more, and that diet and environment and opportunity are all in there too.

That sentence is the 2026 scientific position, printed thirteen years early. Without it the chapter would be judged on its headline, and the headline would fail.

There is one more thing, and it is the reason this book exists at all.

In his 2026 book, writing about how science corrects itself, David reports that in genetics "an entire field ... was basically wiped out," because researchers had been trawling enormous databases for spurious correlations between some human trait and some spot of DNA. He is right, and it is the same collapse described above. The endnote attached to that sentence adds, in a parenthesis: *this was also a topic in my first book, The Sports Gene*.⁴

He described the destruction of the research programme that a substantial part of his first book stood on. He cited his own book as having covered the topic. He did not say what the destruction did to it.

I do not think that is dishonesty, and it is worth being clear about why. He was writing about preregistration, in a chapter about research methods, and the fact was doing a different job in a different argument. Nothing obliged him to stop and audit himself. But it is the strangest thing in three books, and once you have seen it you cannot read the first one the same way.

One last piece of arithmetic, from the book's own epilogue, which is the best answer it gives to its own title. Take the couple of dozen genetic variants that had been linked to endurance performance. Work out the odds of one person holding the ideal version of all of them. It comes to less than one in a quadrillion, and nobody alive is thought to hold more than about two thirds of the set.

Which is the honest ending. There is such a thing as a sports gene, and you will almost certainly never meet anyone who has one. What the rest of us have is thousands of variants each worth almost nothing, in combinations nobody can read, plus a childhood, plus a lake to ski across.

DOES THE SAME TRAINING WORK ON EVERYONE?

David ran middle distance at Columbia, and the best story in his first book is about himself losing to somebody for two years and being wrong about why.

His training partner was faster than him. Faster by five seconds over eight hundred metres, which at that level is not a gap, it is a different category of runner. Both his parents had been university track champions. He held a national age-group record. He was, on paper and on the track, the more gifted man by a distance.

And he stopped improving. Season after season, the same times.

Before one indoor race he announced to the team that he was going to drop out at the six-hundred-metre mark, to relieve himself of the pressure of failing to improve again. He did not drop out. He finished, and David passed him in the last hundred and fifty metres and beat him for the first time.

The conclusion was obvious and everyone on the team had already reached it. The gifted one was soft. He had been handed everything, had never needed to suffer for it, and when the training got hard enough to require suffering he had nothing to reach for. Whereas David, who had less, had made himself into a runner through sheer application.

That is a satisfying story. It is the story every team tells about its talented underachiever, and it is very close to the only story available, because the alternative explanation is invisible.

Then David went looking for somewhere private to throw up after a hard session, and found his training partner already there, retching over a bin. And then again. And then discovered he sometimes darted off in the middle of an interval workout to vomit and came back to finish the session.

He was, in David's phrase, as tough as titanium screws.

So the narrative was wrong. Not shaded, wrong. The two men were doing the same workouts, on the same track, under the same coach, with the same willingness to hurt themselves, and one of them was getting faster and one of them was not.

Something else was going on, and it was not character.

What the book is arguing

The claim is that **trainability is itself a talent**, separate from how good you are now, and that it is substantially inherited.

The evidence is a study called HERITAGE, run across five universities in the 1990s. Ninety-eight families, four hundred and eighty-one people, none of them athletes. Everyone did exactly the same programme: stationary cycling, three times a week, five months, supervised in a laboratory so that nobody could quietly do more or less than anybody else. Everyone was measured before and after, and everyone gave DNA.

The measurement was aerobic capacity, usually written as VO2max, which is the maximum rate at which your body can take in and use oxygen while you exercise. It is the single most useful number in endurance sport, and for most of a century it has been treated as the thing training improves.

The results, as reported: some people improved their aerobic capacity by more than fifty percent. Some people improved by nothing at all. Same programme, same supervision, same five months.

And the improvements ran in families. Related people responded similarly; unrelated people did not. The headline figure to come out of it is that roughly half of how much you respond to training is determined by whom you are descended from.

If that is true, it reorganises everything. It means that when two people train identically and get different results, the difference may be neither effort nor technique. It means "he isn't working hard enough" is often just wrong. And it means the invisible variable is doing the work that we routinely assign to character, which is why David's story about his training partner is the opening of the chapter rather than a footnote to it.

The book also reports the other end of the same idea. Researchers at York University tested nineteen hundred aspiring Toronto firefighters and found six with no training history at all whose aerobic capacity was already at the level of collegiate distance runners. They had the blood volumes of trained endurance

athletes without having done anything. The phrase for them is the naturally fit six, and their existence is the reason the book insists that where you start and how fast you improve are two separate gifts, and that having both is rare.

Why it made sense to say at the time

Because nobody had built anything like HERITAGE before, and the older literature had a hole in it exactly the shape of this question.

The standard way to test whether an exercise programme works is to take a group of people, measure them, train them, measure them again, and report the average change. That design answers "does this work on average," which is a real question. It cannot answer "does this work on you," because it never reports the spread, and if it does report the spread there is no way to tell how much of it is real.

HERITAGE was designed to look directly at the spread and at whether it ran in families. That is a genuinely good idea, it was expensive, and it produced a number that was novel and quotable. In 2013 there was very little standing against it.

Does it hold up

This is the worst-aged material in the book, and the reason is not that somebody came along and refuted it. It is that the design was never capable of establishing it, and the people best placed to notice said so, quietly, and were ignored.

Start with the arithmetic, because it is the whole thing.

Every measurement is the true value plus some noise. Test the same person's aerobic capacity twice in a week with no training in between and you will get two different numbers. The equipment varies, hydration varies, sleep varies, motivation on the day varies. That variation is not a flaw in the study; it is a property of measuring human beings.

Now put four hundred people through a training programme and look at the spread of their improvements. That spread contains real differences in response, and it also contains all of that noise, and a before-and-after study with nobody in it as a comparison has no way of separating the two.

There is a specific consequence with a name. If you select people because their first measurement was extreme, the noise that made it extreme does not repeat itself

on the second measurement, so their scores drift back toward the middle. This is called regression to the mean, and it will manufacture the appearance of individual response out of nothing at all, with no biology involved.

Here is what happens when somebody checks.

A team went back to the one trainability study from that era that had included a control group, meaning people who were measured twice and did no training in between. The trained group improved by an average of five and a half units. The control group, who did nothing for twenty weeks, changed by about half a unit in the other direction, with a spread *larger* than the trained group's.¹

Take the threshold the field conventionally uses to call somebody a responder and apply it to those untrained controls. **Twenty-three percent of the people who did no training whatsoever would have been classified as positive responders.** On noise.

And for an individual's improvement to carry even a seventy-five percent chance of being a real response rather than measurement error, it would have to be more than double the conventional threshold.

Then somebody did the obvious experiment, which had never been done. Take fourteen men, give them an identical four-week training block, wait three months, and give them the same block again. If responsiveness is a property of a person, the same person should respond the same way twice.

The correlation between a man's first response and his second was 0.40 for aerobic capacity, and 0.11 for how long he could sustain effort before fatiguing.² Those are the numbers of a coin, roughly. A responder label that does not survive being applied to the same man twice was never describing a property of the man.

Now the part that should have ended the argument in 2004.

Claude Bouchard ran HERITAGE. He is the most important figure in the entire field. In 2004 he co-authored a paper that took HERITAGE's own headline spread in training response, subtracted the estimated test-retest error, and found that roughly **half of the famous spread was the measuring equipment.**³

His own group. Twenty-two years ago. The paper is careful and two-sided: it says real individual differences do persist after the correction, which is true and which is the responsible thing to say. But both halves of the argument that is still running today are sitting in one abstract published by the people who created the data, and almost nobody cites it.

And there is one more finding, more recent, which does not deflate the claim so much as relocate it.

"Identical training" was never identical. When you prescribe exercise as a percentage of somebody's maximum heart rate, which is how it has almost always been done, you are not giving two people the same physiological stimulus, because the same percentage of maximum sits at very different places relative to each person's actual metabolic thresholds. For one man it is comfortably aerobic. For another it is close to his limit.

Prescribe training relative to those individual thresholds instead, and the proportion of people who improve by a meaningful amount goes from sixteen percent to sixty-four percent.⁴

Read that again, because it reframes the whole literature. A large part of what was recorded as *individual response variation* was individual *dose* variation. Many of the people labelled non-responders were never given the same stimulus in the first place. They were given the same instruction, which is a different thing.

So where does that leave the claim.

The strong version is gone. Zero-to-fifty-percent is not a measured range of human biological response; it is a measured range of human biological response plus measurement noise plus unequal dosing, and nobody knows the split. The specific figure that half your trainability is inherited has never been re-derived by anybody using a design capable of producing it, and a search across more than a hundred and eighty papers from HERITAGE found a comparator arm in none of them.

The weak version survives and is still interesting: people do differ in how they respond to training, probably substantially, probably with some heritable component, and we currently have no honest way to say how much or to identify who is who in advance.

And the thing the story was actually about survives completely. David's training partner was not soft. Whatever was going on between those two men, the moral explanation was wrong, and the moral explanation is the one everybody reaches for because the alternative cannot be seen from the outside. That part needed no statistics and has not aged a day.

There is a coda in the chapter that is better than any of the numbers.

Jim Ryun was the extreme responder of the century. Cut from his junior high basketball team, twenty-first best runner in his school in his first cross-country race, and then, within about two years, the first high schooler to run a mile under four minutes, and later the world record holder.

Years later, after he quit mid-race at a national championship and refused to run at all for a year, his physiologist kept testing him. His aerobic capacity fell by twenty percent.

Where it landed was exactly the level of the naturally fit six: the untrained Toronto firefighters who had never done anything and were already at the level of college runners.

Then he started training again and got essentially precisely that twenty percent back.

That is what the two gifts look like when you separate them. The greatest responder anyone had ever measured, detrained to nothing, was still standing on a floor most people never reach.

CAN YOU SEE IT COMING?

In 2004 a softball pitcher named Jennie Finch was invited to an exhibition game against a lineup of major league baseball players, and she took them apart.

She struck out Mike Piazza on three pitches. She made another hitter miss so badly he spun all the way round. Albert Pujols, who at the time was the best hitter alive, was startled backwards by the first pitch, spun away from the second, swung and missed three times, and walked off. Then he stopped, turned back, and tipped his cap to her. He said afterwards he did not want to experience that again.

Barry Bonds had publicly promised to hit her hard and asked for national television coverage. When she actually turned up he insisted the cameras be switched off, watched a long series of strikes go past, and started arguing that they were balls. A teammate pointed out that there were twelve umpires standing behind him. He eventually managed a foul tip, after Finch told him in advance what she was going to throw.

Alex Rodriguez, the reigning Most Valuable Player, watched his own catcher fail to catch three of her five warm-up pitches and declined to bat at all. Nobody, he said, is going to make a fool out of me.

The fielders behind Finch had already sat down. One of them was lying in the dirt using second base as a pillow.

Now the important part, which is that she was not throwing harder than they were used to. She was throwing *softer*. Softball is pitched underhand from forty-three feet, and her pitches left her hand in the upper sixties in miles per hour. Because the mound is closer, that gives the batter roughly the same total flight time as a ninety-five mile per hour fastball from a baseball mound.

Same time. Easier speed. And the best hitters on earth could not touch it.

What the book is arguing

To understand why, you need the physics of the thing they normally do.

A major league fastball takes about four hundred milliseconds to reach the plate. It takes seventy-five milliseconds just for the cells in your retina to confirm that a ball is in view, during which the ball travels about ten feet. Simple reaction time,

meaning the gap between a light coming on and you pressing a button, is around two hundred milliseconds for essentially everybody. And the window in which the bat can actually make contact is about five milliseconds wide.

Add those up and hitting a fastball is impossible. The advice to keep your eye on the ball is not merely difficult, it is anatomically unavailable: no human visual system can track a pitch all the way in. A batter could close his eyes halfway and lose nothing.

So the book's answer is that elite hitters are not reacting. They are predicting.

Everything a pitcher does before the ball is released, the arm angle, the wrist, the grip visible for an instant, the arrangement of the whole body, is information. A hitter who has seen tens of thousands of pitches has an enormous library of those patterns and what followed them. He begins his swing on the pattern, not on the ball. He is seeing several hundred milliseconds into the future, which is exactly the amount of time he does not have.

This is called **chunking**, and it was worked out in chess long before anybody applied it to sport. Grandmasters shown a real board position for a few seconds could reconstruct it almost perfectly, while weaker players managed half. The obvious conclusion was that chess masters have extraordinary memories.

Then somebody put the pieces on the board at random, in arrangements no real game could produce. The grandmasters' advantage vanished completely. They were not remembering twenty-five pieces. They were remembering five or six familiar clusters, and when the clusters were destroyed they were ordinary people looking at a board.

The sporting version of that experiment is the occlusion test. Show an athlete a film of an opponent and cut it to black at the moment of contact, or before it, and ask what happens next. Experts are excellent at this and non-experts are not. Put opaque goggles on tennis players that black out at the instant of contact, or blur out a bowler's arm from a cricket batter, and elite performers fall back toward novice level.

Which is precisely what Jennie Finch did to a lineup of All-Stars. She was not a better athlete than Albert Pujols. She was an unfamiliar arrangement. Underhand

delivery, different arm path, different spin, from a distance his library had no entries for. She took away his crystal ball and left him with two hundred milliseconds and a five millisecond window, and that is not enough, and it is not enough for anybody.

The phrase the book uses, which it says came up over and over in interviews with the psychologists who study this, is: **it's software, not hardware.**

Why it made sense to say at the time

Notice where this chapter sits. It is the first chapter of a book called *The Sports Gene*, and it is the strongest case for practice in the entire book. That is deliberate architecture. The chapter builds the anti-genetic argument as well as it can be built, and then the next chapter, with the two high jumpers, takes it apart.

But it is also making a real historical point. This body of work is *why* the deliberate practice framework existed at all. If the thing that separates an expert from a novice is an acquired database rather than superior equipment, then expertise is something you build, and the interesting question becomes how many hours it takes to build it. Chunking is the finding that licensed the ten thousand hours idea. Without it there is no reason to think the hours are the story.

And the chapter states the strong version flatly: no one is born with the anticipatory skills required of an elite athlete. That is one of very few completely unhedged sentences in the book.

Does it hold up

It splits three ways, and the split is instructive because it is the same claim looked at from three distances.

The anticipation finding is vindicated, comprehensively. A meta-analysis of forty-two studies and nearly four hundred effect sizes found experts superior at picking up cues, faster and more accurate, with systematically different eye movements: fewer fixations, held longer.¹ The random-chessboard reversal has never been overturned in any domain anyone has tried it in. Modern brain imaging finds experts performing their own task with *less* prefrontal activation than novices, which is what an automated, acquired skill looks like from the inside. This is among

the most solid findings in sports science and the Finch demonstration is a genuine natural experiment that needed no help.

The reaction time claim is where it gets awkward. The body text says elite athletes do not have faster reflexes than ordinary people, and the evidence offered is that Pujols scored in the sixty-sixth percentile for simple reaction time against a sample of college students.

Two problems. The smaller one is that the sixty-sixth percentile is above average, and it is being used to argue that his reaction time is unremarkable. The larger one is that the broader literature does not agree. Athletes do come out faster than non-athletes on speeded laboratory tasks, with the largest advantages in exactly the interceptive sports the chapter is about.² The effect is not enormous, and the field's own reading is that it lives in decision and attention rather than in raw sensorimotor speed, which leaves the narrow version of the claim standing. But the flat version does not survive.

And here is the thing worth noticing, because it becomes a pattern across all three books. **The correct, hedged version is in a footnote.** Down at the bottom of the page, the book concedes that some rare athletes simply do have superior reaction speeds, and cites Muhammad Ali reacting to a light in a hundred and fifty milliseconds. That is a direct contradiction of the sentence in the body text, parked where a reader will not weigh it against the sentence it contradicts.

The vision claim is the one that has aged worst, and the reason is a comparison that was never like for like. The chapter reports that professional baseball players have extraordinary visual acuity, far outside the general human distribution, and treats this as innate hardware rather than a training effect.

The group difference is real and has been replicated: professional players do measure better than matched controls on static acuity and on dynamic acuity while the head is moving.³ But the phrase *far outside the general human distribution* rests on setting a few hundred selected young American professional athletes, measured on a custom eye chart built to go below the normal floor, against general-population eye studies conducted in India and China across the whole adult age range. Those are not the same kind of sample and the gap between them is not all vision.

And then there is the finding that cuts the thesis in half.

A 2026 meta-analysis pooled twenty-six studies of visual perception in fast-paced ball sport athletes. Elite athletes came out far ahead of non-athletes: a large effect, unambiguous.⁴

Elite athletes came out no better than **sub-elite** athletes at all. The effect was 0.12, and not statistically distinguishable from zero.

Sit with what that means. If basic visual function were the fixed hardware that determines who reaches the top, it ought to separate the top from the nearly-top. That is the comparison that matters, because everyone in it has already been filtered for everything else. It does not separate them. Whatever is deciding who makes the major leagues rather than the minors, this is not it.

The same analysis found the advantage was larger for *perceptual* function than for basic *visual* function, which is to say: the acuity end of the range, which is the chapter's evidence, is the weaker end. And separate work shows visual-perceptual performance is trainable to a moderate degree, which dissolves the inference the chapter was making, because that inference was "these athletes are outside the normal range, therefore it cannot be a training effect."

So the honest version of the chapter, thirteen years on:

Anticipation is learned, it is sport-specific, it evaporates the moment you change the cues, and Jennie Finch proved it in front of a crowd. That stands and is not in any trouble.

Elite athletes are also modestly faster and modestly sharper than the rest of us in the laboratory, which the chapter denies in the body and concedes in a footnote.

And the difference between an outstanding professional and an ordinary one is not in the eyes.

WHY DO THE BODIES KEEP GETTING STRANGER?

If you look at photographs of athletes from the first half of the twentieth century, the striking thing is how ordinary they are. A shot putter and a high jumper and a swimmer from 1925 look like three men. Big men, fit men, but recognisably the same animal.

Now look at a modern Olympics. The gymnasts and the basketball players are not the same animal. The distance runners and the shot putters are not the same animal. Elite volleyball players and elite coxswains barely overlap on any dimension you could measure. Somewhere in between those two sets of photographs, the bodies came apart.

The book calls this the Big Bang of body types, and the metaphor is doing real work: a period during which everything was clustered together, followed by a rapid expansion outward in every direction at once.

What the book is arguing

The claim has two halves.

The first is that this actually happened, measurably, and it is not just an impression from photographs. Two Australian scientists compiled body measurements of elite athletes across the twentieth century and found that in sport after sport, athletes were getting taller, or heavier, or longer-limbed, and pushing further from the average in whatever direction their sport rewarded.¹

Now, populations in general got taller during the twentieth century, because of nutrition and public health. That is called a secular trend and it would produce bigger athletes all by itself, without anything interesting going on.

The finding that matters is that the athletes moved faster than the population did. The rates of change in elite sport outstripped the secular trend. Whatever was happening was not simply everybody growing. It was a filter getting narrower.

The second half is the explanation, and it is economics rather than biology. Over that same century, sport went from something people did in their own town for a trophy to a global market in which a small number of people are paid extraordinary

sums. When the prize is that large and the search is that wide, every sport can afford to become monstrously specific about what it wants. It stops recruiting from the local pool of decent athletes and starts scanning the planet for the one body shape that suits it best.

Which produces a second, more unsettling claim. If you compute the fraction of the general population whose height and weight fall inside the range occupied by elite performers in a given sport, the numbers are tiny. Something like a quarter of people are in the zone for soccer or sprinting. Fewer than one in ten for rugby forwards. The comparison the book offers is fashion modelling, where the fraction who fit the international standard is well under one percent.

The uncomfortable version of that: for most elite sports, most people are not in the running before anybody has assessed a single thing about their character, work rate or skill.

Why it made sense to say at the time

Because the underlying dataset was real, the analysis was published, and the mechanism is straightforward once someone points it out.

And because it explained something the ten-thousand-hours framework could not. If expertise were mostly accumulated practice, the physical shape of the people accumulating it should not have changed systematically over a century. It did, and in directions specific to each sport, and faster than the population underneath it. That is a hard fact for a purely environmental account of who succeeds.

It also let the book do something more interesting than assert that genes matter. Nobody's genes changed over that century. What changed was the *sorting*. The same distribution of human bodies existed in 1925; the market simply got much better at finding the ends of it and much more willing to pay for them.

Does it hold up

The historical claim holds. The forecast does not, and the forecast is the most instructive thing in the chapter.

On the divergence itself: solid, well-cited, and the clause that carries the whole argument, that athletes outpaced the secular trend, is the original researchers' own finding rather than an interpretation laid on top of it.

On the argument with Gladwell, which the chapter is partly having: *Outliers* proposed that height in basketball works as a threshold. Past about six foot, the idea goes, extra inches stop mattering much and other things take over.

The book says no: the probability of reaching the NBA keeps climbing steeply with every additional inch, with no plateau anywhere you can observe. That direction is not in dispute anywhere in the anthropometry literature and it is the sentence that actually refutes the threshold idea.

The famous endpoint attached to it, that something like seventeen percent of American men over seven feet tall are in the NBA, is a figure David modelled himself, and it depends heavily on an estimate of how many seven-footers exist, which is not well established. It is an arresting number. It is also the one part of the argument that does not need to be true for the argument to work, and I would not repeat it without checking.

The related claim about arm span holds up better than it might sound. NBA players really are systematically longer-armed than their height predicts, by something like four to six percent, against a general population where arm span and height are roughly equal. That is far enough from normal that a similar ratio is used clinically as a screening flag for a connective tissue disorder. The precise figure in the book is again David's own analysis rather than a published one, and he says so.

On the muscle ceiling: the chapter reports that the skeleton limits how much muscle a body can carry, at roughly five kilograms of muscle per kilogram of bone, attributed to the measurements of a kinanthropometrist named Francis Holway across thousands of elite athletes.

I could not find that number published anywhere. Holway is real, active and well published, and the ratio does not appear in any of it. The qualitative point is uncontroversial: skeletal size does constrain sustainable muscle mass, and practitioners do use anthropometry to set individual targets. But the specific figure

rests on an interview, and the corroborating evidence offered in the book is Holway's experiment on himself.

And then the forecast.

The chapter observes that world records had largely stopped falling. Records in many events had been broken repeatedly through the 1970s, 80s and 90s, and then, around the turn of the century, the progression flattened out.

The explanation offered is that the Big Bang has run its course. The global search has now found the rare bodies. There is no undiscovered population of enormous or tiny or long-limbed people left to recruit, so the physical raw material is exhausted and the records will not move much again.

That was a real observation about a real plateau. And then the plateau ended.

Distance running records fell comprehensively over the following decade, and the field credits footwear. Advanced racing shoes introduced from 2017 improve running economy by a few percent, and a few percent is enormous. The modelling paper that documents the reversal says so plainly: after two decades of stagnation, recent improvements enabled new records, aided by recent technological advances.²

What makes this the best teaching case in three books is not that he got it wrong. It is *how*.

He saw the plateau correctly. He offered a mechanism that could not be tested, because there is no way to measure the world's remaining supply of undiscovered rare bodies. And the actual cause was something he had already named in the same chapter and set aside as a lesser factor.

Observing a plateau and explaining one are different acts, and only the first was well supported. When a trend is flat, the flatness is data. Every story about *why* it is flat is a hypothesis, and the ones that cannot be tested should be held loosely, especially when they are load-bearing for a prediction about the future.

The bodies did diverge. The market did get better at finding the extremes. What did not happen is the ending, where the search runs out of people and human

performance settles down. It ran out of shoes instead, and then somebody made better shoes.

DOES ANCESTRY EXPLAIN WHO WINS?

There is a region in Kenya's Rift Valley, home to a group of people called the Kalenjin, which has produced a share of the world's elite distance runners so disproportionate that it does not look like a statistic. It looks like a mistake.

Jamaica, an island of under three million people, has spent two decades doing something similar in the sprints.

Everybody notices. And the noticing produces two immediate and opposite reactions, both of which are about the observer rather than the evidence: it is obviously genetic, or it is obviously not, and how dare you ask.

The reason this chapter is worth reading is that somebody went and did the studies, and the answers are more specific and more interesting than either reaction.

What the book is arguing

The book's position on East African running is, unexpectedly, mostly negative. It spends the chapter dismantling the popular biological explanations, one at a time, using research done by a Danish group in Kenya in the late 1990s.

Do Kalenjin runners have more slow-twitch muscle fibre, the endurance-favouring kind? No. Muscle biopsies found no difference from European runners.

Do they have higher natural aerobic capacity? No. Elite Kenyan and elite European runners reach similar values, and there is no sign that untrained Kenyans start higher than untrained Europeans.

Do they respond faster to training? No evidence of it.

Three plausible mechanisms, three nulls, all from the same research programme, and the book reports all three.

What the Danish group *did* find was about legs. They compared untrained Kalenjin boys with untrained Danish boys and found the Kenyans had markedly thinner lower legs, and better running economy, meaning they used less oxygen to run at a given speed.

The mechanism is straightforward physics. Weight at the end of a limb costs far more energy to move than weight near the body, because you are swinging it

through a longer arc thousands of times per hour. A kilogram at the ankle is worth several at the hip. Thin lower legs are, for a runner, free money.

The book is careful about what that does and does not mean. Lithe legs help running economy regardless of who has them. And running economy is separately shown to be trainable, so nobody should read this as a fixed ceiling.

Alongside that, two non-biological arguments.

The first is that Kenya has an extraordinary talent sieve. A very large number of children run substantial distances daily, often to school and back, in a place where success in running is one of very few available routes to money. When researchers surveyed elite Kenyan runners, twenty-two percent of ordinary controls had run to school as children, seventy-three percent of national-class athletes had, and eighty-one percent of international-class athletes had.¹ That gradient is the evidence, and it is more persuasive than any single number.

The second is that part of East African dominance may not be East Africans getting better. It may be everybody else getting worse. The book documents the collapse: American men running under two hours twenty in the marathon fell from 267 in 1983 to 35 in 1998. Britain fell from 137 to 17. The United States qualified one man for the 2000 Olympic marathon; Finland, which once owned distance running, qualified nobody in any distance event.

The line the book quotes, from a coach, is the best sentence in the chapter: the genes did not go away in Finland. The culture did.

On Jamaica, the book examines and rejects a specific folk theory: that the slave trade selected for physical hardiness, so that Caribbean populations descend disproportionately from those who survived a brutal filter, and that this shows up in sprinting.

Why it made sense to say at the time

Because it was, and still is, the only body of physiological work anybody has done on the question, and because the alternative to reporting it was to leave the field to people making assertions.

There is also an argument in the chapter that deserves more credit than it gets. People are fond of computing per-capita medal ratios and presenting them as though a large number implies biology. The book points out that this reasoning is worthless, because a per-capita ratio conflates population size with participation, funding, coaching, facilities, opportunity cost and the simple fact that different countries care about different sports. German dressage and Canadian hockey produce equally striking ratios and nobody reaches for a gene.

That argument is correct, and it is worth noticing that it cuts both ways. An impressive medal count is not evidence for a genetic explanation and it is not evidence for a cultural one either. A statistic that cannot distinguish between two hypotheses is not evidence for whichever one you prefer.

Does it hold up

Better than most of the book, and the reason is that where the evidence was thin the book said so, and where it refused to conclude, the refusal turned out to be right.

The three nulls are still the field's position. The most recent authoritative review is by the same Copenhagen investigator who did the original work, and it restates all three: similar maximal oxygen uptake, similar fractional utilisation, no difference in slow-twitch fibre proportion.²

Which brings a fact worth sitting with. The central empirical foundation of the world's most-discussed question about ethnicity and athletic performance is a single research programme from the late 1990s that nobody has repeated in thirty years, and the field's most recent review of it is by the man who ran it, ending with a request for the studies that would settle the matter. The book reported that honestly at the time. It is still true.

The lower-leg finding has been strengthened, and in the most useful possible way. The same group re-ran the protocol on a larger scale, comparing elite and sub-elite runners from both countries plus untrained adolescents from both.³ The oxygen cost of running was substantially lower in the Kenyan runners, and the lower-leg difference showed up in the untrained adolescents too, which is what rules out the

obvious objection that thin calves are simply what you get from running a great deal.

This is the best claim in the chapter and the model for how the whole subject should be handled. It is quantitative. It has a stated physical mechanism. It survived the obvious confound. And it is a claim about a body part rather than about a people.

The caveat to state plainly: it is the same research group extending its own programme, not an independent replication, and nobody outside Copenhagen has tested it.

The Jamaican genetics came out against the folk theory, twice over.

The research team that had been collecting Jamaican sprinter DNA published its results. Comparing elite Jamaican athletes against Jamaican controls, they found no significant difference in maternal lineages, and concluded that the sprinters come from the same source population as everybody else on the island, with no sign of stratification or isolation for sprinting.⁴ Later genome-wide work found no loci associated with athlete status in Jamaicans either.

And the selection hypothesis, the idea that the Middle Passage acted as a filter whose effects are still detectable, was tested directly. A genome-wide scan of 29,141 African Americans looked for signatures of directional selection since admixture and found none.⁵

Then there is the finding nobody talks about, and it runs the other way. In the same Jamaican study, the researchers also examined elite African-American sprinters. There, they did find something significant: elite sprinters carried maternal lineages of non-sub-Saharan origin at twenty-one percent, against eight percent in controls.

More European maternal ancestry among the elite sprinters, not less. It is the only statistically significant ancestry-and-athlete association anybody has published, it points in the opposite direction to the folk theory, and in the years since, nobody has followed it up.

The sickle cell material has been partly reversed, and this one matters because it touches medicine rather than sport.

The book reports the malaria-adaptation story, which is solid: the sickle cell trait persists in populations from malarial regions because carrying one copy protects against malaria. It also reports, as established, that sickle cell trait can cause sudden death during extreme exertion, an attribution that had led American college sport to mandate screening.

In 2025 the American Society of Hematology published a systematic review. It screened nearly fifteen hundred citations, found seven studies with usable original data, and concluded that sickle cell trait alone does not cause exertion-related death, and that recording it as a cause of death is medically inaccurate.⁶ The screening mandate it had justified remained in place.

The chapter also relays a further hypothesis, that reduced oxygen-carrying capacity in West African populations drove a compensatory shift toward fast-twitch muscle. The book labels it explicitly as unproven and untested in humans.

Thirteen years later it has still never been tested in humans. Not once. The refusal to endorse it was correct, and the reason it was correct is not that the idea was disproven but that nobody has done the work, which is a different and more uncomfortable situation.

The altitude material is the best science in the chapter and it got better.

Three populations have lived at high altitude for thousands of years, and all three solved the oxygen problem differently. Andeans increased red blood cell concentration. Tibetans did not, and instead breathe faster with wider blood vessels. Ethiopians did something else again that is still not fully understood.

A year after the book was published, researchers established where the key Tibetan variant came from. It was inherited from Denisovans, an archaic human population, through interbreeding tens of thousands of years ago.

That is a spectacular result, and it carries a lesson pointing the opposite way from the rest of this chapter. The Tibetan case is clean precisely because it is a large-effect variant, in a well-defined biological pathway, with an unambiguous physiological readout. That combination is exactly what athletic performance genetics has spent forty years failing to produce. When the biology is genuinely

there and genuinely large, it does not require sophisticated statistics to see. You can find it with a few hundred people and a blood test.

And on the West getting slower, time ran an experiment nobody designed.

The book's argument in 2013 was that the rich world had become richer, heavier and less interested, and that this accounted for part of East African dominance.

The historical trend was real and correctly reported. But the rich world then got considerably faster again. American marathon depth recovered enormously: the United States qualified a single man for the 2000 Olympic marathon, and had well over two hundred men under two hours eighteen for its 2024 trials.

East African dominance did not move.

Nobody has formally compared the eras, so this is not a published refutation. But a plausible partial explanation lost most of its force without anybody publishing a word about it, and that is worth seeing.

What the chapter actually establishes

Strip it down and the surviving list is short, specific, and almost entirely about mechanism rather than about people.

Thin lower legs reduce the energy cost of running, this difference is present in untrained adolescents, and it is one real physiological finding in a field of nulls.

Three popular biological explanations for East African dominance have been tested and none of them holds. Kenya's talent sieve is enormous and the running-to-school gradient is steep. Jamaican sprinters are genetically indistinguishable from other Jamaicans. The slave-trade selection hypothesis has been tested at scale and failed. The one significant ancestry finding anybody has published points the opposite way from the folk theory and has never been followed up. And a claim about muscle fibre that the book explicitly declined to endorse has, thirteen years on, still never been tested in a human being.

That last one is the pattern worth carrying forward. Across this whole subject, the places where he refused to draw a conclusion are the places that have aged best, and it is not because he was lucky. It is because the evidence was not there, he said so, and it still is not there.

WHO COUNTS AS A WOMAN?

In 1985, a Spanish hurdler named María José Martínez-Patiño arrived at a competition in Kobe and realised she had forgotten her certificate of femininity.

That was a real document. Female athletes carried one. If you did not have it with you, you took the test again: a swab of cheek cells, examined for chromosomes.

Two months later she was told there was a problem. All fifty cells they analysed had XY chromosomes.

She had never known. She had been raised as a girl, had a female body, and had a condition called complete androgen insensitivity, which means the cells cannot respond to testosterone at all. Her body had ignored every androgen it ever produced.

Officials suggested she fake an injury and retire quietly. She refused, and three months later won the Spanish national title in her event, and the result leaked.

Everything that could be taken was taken. The title, her place in the athletes' residence, her scholarship, her records. Her fiancé left. She was reinstated three years later after a geneticist took up her case, and she missed qualifying for the 1992 Olympics by one tenth of a second.

Five years after her case, the international governing body convened a panel of experts to settle, once and for all, how to tell a man from a woman for the purposes of competition.

The panel's recommendation was to stop trying.

What the book is arguing

The chapter makes two claims that people usually assume are opposites, and it insists on both.

The first: human biology does not sort into two clean boxes, and no test can make it. Chromosomes do not do it, as Martínez-Patiño demonstrates: XY, and no ability to use testosterone. Anatomy does not do it. Hormone levels do not do it, because they overlap and because some bodies cannot use what they produce. There are

numerous ways for the developmental cascade to go somewhere other than the two common destinations, and the resulting people are not rare enough to ignore.

The book states this as strongly as it states anything. No technological advance of the previous two decades had made the slightest difference to the problem, and none would in the future.

That last clause is a prediction, and we can now check it.

The second: the athletic gap between elite men and elite women is real, large, and biological. In running, across every distance from the sprints to the ultramarathon, the top men are about eleven percent faster than the top women. In the long jump the gap is around nineteen percent. In the eight hundred metre freestyle, which is the narrowest gap in any major event, women are within about six percent.

To make it concrete: the women's hundred metre world record would not have been fast enough to enter the men's field at the 2012 Olympics. The women's ten thousand metre world record holder would be lapped by a man running the minimum qualifying standard.

The book is arguing against a specific and once-popular position. In the 1990s and 2000s several papers in serious journals extrapolated the improvement curves of men's and women's records and concluded that women would overtake men, one of them putting a date on it. A survey before the 1996 Olympics found two thirds of Americans expected top women to beat top men eventually.

The book says those extrapolations were wrong, and offers a reason the convergence looked real at the time: a large part of the rapid improvement in women's records through the 1970s and 80s came from systematic state doping programmes, most thoroughly documented in East Germany. When that stopped, the improvement stopped, and some of those records still stand.

The explanation for the gap is sexual selection: in a species where males competed physically with each other, traits useful in fighting were amplified in males, and puberty amplifies them further.

And there is a coda that gives the chapter its title. Women are as athletic as they are, the book suggests, partly *because* men are. The sexes share nearly all of their

genes and cannot be programmed independently, so an adaptation selected in one shows up partially in the other. Men have nipples for the same reason.

Why it made sense to say at the time

In 2013 chromosomal sex testing had been abandoned, and the Semenya case of 2009 had made the subject radioactive without resolving anything. The predictions about women overtaking men were still in circulation and had never been properly answered in a popular book.

And the two halves of the argument were, and are, genuinely uncomfortable together. If you say the categories are biologically messy, you are assumed to be arguing that the athletic gap is a social artifact. If you say the gap is large and biological, you are assumed to be arguing that the categories are simple. The chapter refuses both inferences, which was the right call and the hard one.

Does it hold up

The prediction is the strongest forward-looking call in the book, and it came in.

Since 2013 the regulatory world has worked through the entire menu. Testosterone thresholds were introduced, revised, litigated and revised again, keyed to circulating hormone levels. When those proved fragile in court and in the science, the authorities went back to genetics.

World Athletics adopted a once-in-a-lifetime SRY gene test from September 2025. World Boxing adopted mandatory PCR sex testing from July 2025, extended to all adult athletes from the start of 2026. The International Olympic Committee adopted the same screen in March 2026.

None of it resolved the classification problem, and the professional genetics bodies said so within weeks, in their own voices. The European Society of Human Genetics stated that such a test detects a proxy rather than the thing of interest. Equivalent bodies elsewhere said the same. So did one of the scientists who co-discovered the SRY gene in the first place.

And here is the detail that settles it. Both policies had to write in an exemption for complete androgen insensitivity, which is Martínez-Patiño's condition. A policy

whose entire premise is that one gene answers the question begins by listing the cases where the gene does not answer the question.

The exception is the admission.

A necessary boundary, and I want to be exact about it. What was predicted is that no *test* would classify cleanly. That is a claim about biology and measurement, and it has been repeatedly confirmed. It is not a claim that sport should not have a women's category, and the book does not make that claim. Whether to have a protected category, and where to draw its boundary, is a question about what a competition is for. Evidence constrains that question. It does not answer it, and no amount of additional science will.

The performance gap numbers were right, and one word was wrong.

The quantified claims hold up almost exactly. The most relevant analysis found the gap stabilised after 1983 at about ten percent across eighty-two events, with running around eleven, jumps around eighteen, swimming around nine.¹ Those are the book's numbers to within a point. The most recent consensus statements give ten to thirty percent depending on event, and the picture has not moved.

But the source says *stabilised*. The book says *widening*. That is one step beyond its own best evidence, in a paragraph where every number was excellent, and it is the sentence that was written to be quoted.

There is also a genuinely new fact that nobody had in 2013 and which makes the point far better than "widening" ever did. Analyses now covering the whole lifespan find the sex gap in running roughly *triples* between age thirty and age eighty. The gap is not a fixed property of the sexes. It is a moving thing that tracks the underlying biology through life.

The doping explanation for the 1970s convergence holds. The documentary record from the East German programme has only strengthened, and it makes the book's point unprompted: the programme's own papers record that androgens were emphasised for women and adolescent girls specifically, because the effect there was largest.

One claim needs a surgical correction, and the correction is more interesting than the original.

The book reports that XY women with androgen insensitivity are markedly overrepresented among elite female athletes, on the strength of endocrinologists' impressions, which it flags as impressions.

The overrepresentation is confirmed and quantified: screening of elite female athletes found the prevalence of hyperandrogenic XY differences of sex development at roughly a hundred and forty times the general population rate.²

But the condition is the other one. Among elite female athletes with XY differences identified in a 2026 analysis of nearly seventeen hundred athletes, twelve had 5-alpha-reductase deficiency and two had androgen insensitivity, only one of those complete.³

That reversal matters, because the two conditions are opposites in the relevant respect. Androgen insensitivity means the body cannot use testosterone. 5-alpha-reductase deficiency means it can. So the XY athletes who reach elite level are overwhelmingly the ones whose bodies *do* respond to androgens.

Which is evidence for the testosterone account, and against the alternative the book floated, that something about the Y chromosome other than testosterone might be doing the work. He hedged toward the wrong alternative, and because he hedged rather than asserted, it costs him nothing.

And the most delicate claim in the chapter has aged remarkably, for a reason worth naming.

The book reports a self-experiment. A medical physicist named Joanna Harper, who had transitioned, kept careful data on her own running and found she slowed within about a month of beginning hormone therapy, faster than she expected. Her line was that she felt the same when she ran, she just could not go as fast.

That is an n of one, in a magazine anecdote, and the book presents it as exactly that, with five corroborating cases and no policy conclusion drawn.

Harper subsequently became a researcher. Her 2025 study of ten transgender women runners found a decline of about fifteen percent over roughly two and a half years.⁴ The anecdote's subject published the study that reproduced the anecdote, with ten times the sample and measured hormone levels.

Now the harder thing, and it is the reason this passage aged well.

The book's claim is that suppression produces a large decline *relative to the athlete's former self*. That is what the data show, and it has held.

The question the world has spent thirteen years fighting about is a different one: how such an athlete compares to other women. The book does not answer it. It declines to.

Everything currently contested is downstream of a comparison the 2013 passage did not make. And the samples in that fight remain tiny: eleven, fifteen, ten, twenty-three. When you see a large number attached to this literature, check what it is counting, because the biggest studies are large precisely because they are mostly not about athletes.

The evolutionary coda is untested and should be read as an idea rather than a result.

The underlying mechanism is real and well developed. Males and females do share nearly all of their genome, a trait cannot be optimised independently in the two sexes, and this constraint is actively researched under its own names. What is not established is any step of the specific inference the book draws about athleticism, and recent modelling suggests the intuition linking physical dimorphism to shared genetic architecture does not follow as cleanly as everyone assumed.

The nipple analogy is doing rhetorical work the genetics does not yet underwrite. Which is fine, and the book hedges it properly, attributing it and using the word *might*. But it is an illuminating way of thinking, not a finding.

One number I could not check. The chapter reports that the sex difference in throwing velocity is three standard deviations, twice the disparity in height, such that 997 men out of a thousand would out-throw the average woman. The primary source is old and I could not obtain its full text to confirm the figure. The direction is not in doubt and is supported by other evidence in the same passage. The specific number is unverified here.

What survives

The categories are biologically messy, no test resolves them, and the people running sport have now spent thirteen years and three testing regimes demonstrating it.

The performance gap is real, large, stable rather than widening, and varies by event and by age in ways that track biology. The convergence of the 1970s was substantially pharmaceutical. XY differences of sex development are enormously overrepresented at the elite level, and the ones that dominate are the kind that can use androgens rather than the kind that cannot.

And on the question everybody actually argues about, the book's position thirteen years on is that it did not take one.

That was not evasion. It was the correct reading of what the evidence could support, and it is why this is the chapter that has aged best.

WHAT IS GENETIC TESTING ACTUALLY FOR?

A cardiologist in the book describes telling a seventeen-year-old American football player that he has hypertrophic cardiomyopathy, a thickening of the heart muscle that can cause sudden death during exertion, and that he is finished with competitive sport.

It is a terrible conversation to have with a teenager and the cardiologist does not pretend otherwise.

The book's response to it is one sentence, and it is the emotional hinge of the chapter: it is better, though, than the linebacker himself being gone.

Hold on to that sentence. We are going to come back to it, because it is no longer true in the way it was true when it was written, and the reason it stopped being true is the best news in this chapter.

What the book is arguing

The argument is a division of labour, and it is unusually clean.

Genetic testing is worthless for predicting athletic talent. The chapter says so flatly. The commercial tests then being sold to parents of eight-year-olds were, in the book's phrase, nearly worthless. A stopwatch outperforms them.

Genetic testing is genuinely useful for predicting risk. And that, the book argues, is where the real, present-day, non-hypothetical value of genetics in sport already sits.

Four examples are offered.

Hypertrophic cardiomyopathy, the linebacker's condition, which the book calls the most common cause of sudden death in young athletes. It is genetically diverse, often specific to a single family, and it sits in an awkward diagnostic zone where an ultrasound of a very fit young heart can look worryingly like a diseased one, because endurance training thickens heart muscle too. A genetic test can resolve the ambiguity that imaging leaves.

ApoE4, a gene variant associated with worse recovery from brain trauma, and overrepresented among boxers and footballers with lasting damage. The book

reports this carefully: the association looks real, the extra risk cannot be quantified, most doctors oppose offering the test, and most athletes say they want it.

Connective tissue genes associated with tendon and ligament injury risk.

Pain genes, several variants associated with individual differences in pain perception.

Why it made sense to say at the time

Because the contrast was genuinely stark and nobody was drawing it.

The talent-testing industry in 2013 was selling parents a cheek swab that would tell them whether their child was a sprinter or an endurance athlete, on the basis of one or two variants of trivial effect. That was straightforwardly junk, and saying so plainly was useful.

Meanwhile the cardiac work was real medicine. Sudden cardiac death in a young athlete is rare, catastrophic and heavily reported, and there was a genuine question about whether genetic testing could do what echocardiography could not.

Setting the two side by side made a point that is easy to miss: the problem with consumer genetics was never that DNA is uninformative. It was that predicting a complex trait built from thousands of tiny effects is a fundamentally different task from detecting a single variant with a large effect on one protein. One of those problems was solvable in 2013 and the other was not.

Does it hold up

The division still holds. Testing is useless for talent and useful for risk.

But it holds for **one of the four reasons given**, and the one that held has changed underneath in three ways that matter more than the verdict.

The cardiac case held, and did better than the book predicted.

Genetic testing of a patient with a confirmed diagnosis is now a top-tier guideline recommendation in the United States, and the reason is exactly the one the book identified: it enables cascade screening. Establish the variant in one person and you can test their relatives directly for that variant, converting a single diagnosis into a family's worth of decisions.¹ The book also anticipated, without naming it, the

practice that most clearly saves lives here: molecular autopsy after a young athlete dies, followed by testing of surviving relatives.

But the genetics got smaller, not bigger.

The book reports eighteen genes and 1,452 known mutations, "and counting," which paints a picture of a field expanding into ever greater complexity.

What actually happened was curation. An expert panel systematically reappraised the genes said to cause the condition, and in its 2025 reassessment changed seventeen of thirty-one classifications it revisited and **downgraded nine genes to "disputed" specifically to discourage clinical laboratories from reporting them.**²

The field ended with about twenty-nine genes carrying real evidence, and a core of nine doing most of the work.

That is what a maturing field looks like from the inside. It does not accumulate. It prunes.

And variant interpretation turned out to have an error rate, which is now measured.

This is the single most useful thing anybody can tell you about clinical genetic testing in 2026.

When a large patient registry systematically re-examined the variants it had on file, seventeen percent were reclassified. That included **108 variants downgraded from "pathogenic" to "uncertain significance," affecting 196 patients**, and sixty-one moved the other way, affecting a further 199.³ A single Italian centre re-examining its own records reclassified twenty-two percent of its variants.

People had been told they carried a disease-causing mutation, and later they had not.

A pathogenic call is a snapshot of what was known when it was made, not a permanent fact about your body, and periodic reinterpretation is now a recognised clinical duty rather than an optional courtesy. The book's hedge, that this is an inexact science, was right. What it could not have known is how inexact, or that somebody would eventually put a number on it.

There is a redeeming detail. At the Italian centre, cleaning up the classifications made the remaining genotype *more* prognostic: after reclassification, the association between carrying a variant and actual bad outcomes got stronger. Removing the wrong calls made the right ones mean more.

Now the reversal, and it is the reason this chapter needed writing carefully.

The linebacker sentence reflected the standard of care in 2013, which was blanket disqualification from competitive sport on diagnosis. Given that, "better than the linebacker being gone" was the correct and humane thing to say.

It is no longer the standard of care.

A 2025 scientific statement from the American cardiology bodies organises the entire question of sports participation with cardiovascular abnormalities around **shared decision-making**.⁴ A 2026 review in the *New England Journal of Medicine* states the position directly: a growing body of evidence indicates that many athletes can return to play after disease-specific treatment without increased risk, and professional societies now consider return to sport reasonable through shared decision-making.⁵

The seventeen-year-old in 2026 does not get an automatic ban. He gets a conversation, an individualised risk assessment, an emergency action plan, and a defibrillator on the sideline.

Which means the emotional argument of the chapter has been overtaken by good news. The trade-off it dramatised, a career against a life, is much less often the actual trade-off. Anybody reprinting that exchange today without saying so would be reproducing a superseded standard of care in a book about medical testing, and that is worth more than a footnote.

The epidemiology also moved. "The most common cause of sudden death in young athletes" was defensible on the registry data available in 2013. Current work attributes a substantial minority of cases to cardiomyopathies, with this condition the largest single identifiable one, while a large group of cases show a structurally normal heart at autopsy. The phrase that fits now is *one of the leading identifiable causes, alongside a larger group where the autopsy finds nothing*.

Two of the four examples have no clinical standing at all.

Connective tissue genotyping for injury risk is recommended by no professional body. The researcher quoted in the book supplied his own limit at the time, saying that all they could tell an athlete is that a particular profile carries increased risk based on current knowledge. That is still the whole of the case and it has not strengthened in thirteen years.

Pain genetics is not a clinical test in sport in any form. One of the genes named is a rare-disease diagnostic used for something else entirely. Another is a curiosity with contradictory findings. The third survived meta-analysis only as a small effect in chronic pain patients and returned null in a large study of cancer pain.

ApoE4 is the instructive middle case, and it is instructive because everything the book said about it is still true and the ground moved anyway.

The book reported four things: the association looks real, the increment cannot be quantified, the doctors are opposed, the athletes want it. All four remain accurate descriptions of the situation.

What changed is that the association is not where the early boxing and football studies pointed. A 2026 systematic review across thirty-three studies reports **no relationship between the gene and concussion risk.**⁶ The consistent signals for that variant are elsewhere: in bleeding within the brain, and in the degenerative disease found at autopsy in some former players.

And the thing athletes actually wanted, which was a way to know something about their own brains, arrived. It just did not arrive as a genotype. It arrived as blood tests measuring proteins associated with brain injury, which are far more informative than a gene variant because they measure what has happened rather than what might.

The ethical worry raised in the book, about who gets to see the result, did not go away. It transferred to a better instrument, which makes it worse. The book correctly notes that American genetic non-discrimination law does not cover life, disability or long-term care insurance. It also does not cover a blood test for a protein, because that is not genetic information at all.

The consolidated answer

Yes, the division still holds. Genetic testing remains useless for predicting talent and genuinely useful for predicting risk.

But the useful part is narrower and deeper than the four examples suggest. It is essentially one domain, inherited cardiac disease, where testing is guideline-recommended, aimed at families rather than individuals, subject to mandatory reinterpretation because a meaningful fraction of past calls were wrong, and no longer a sentence ending a career.

Everything else on the list has either stalled where it was or migrated out of genetics entirely.

And the sentence about the linebacker, which was true and humane when written, is the thing in three books that has been most thoroughly overtaken by the world simply getting better at something.

DID THE TEN-THOUSAND-HOURS ARGUMENT EVER GET SETTLED?

The book opens with two high jumpers.

One was a Swede who had been training since he was four. Twelve sessions a week. He would not leave practice on a miss, and it is a fair bet he has taken more high jumps than any human being who has ever lived. It got him an Olympic gold medal.

The other was a college basketball player who had never high jumped in his life. Somebody dared him onto a track. On the seventh jump he ever took, he cleared a height most jumpers chase for a decade.

Twenty months later the two of them met at the world championships, and the basketball player won.

When scientists eventually looked at both men, the difference came down mostly to one tendon. The Swede's Achilles had been made extraordinarily stiff by twenty years of landing on it. The basketball player's was simply very long for his height, and a long one does the same job for free.

One of them built the machine. The other was issued it.

What the book is arguing

The target is the idea everybody had absorbed by then, which arrived by a route worth knowing. A Swedish researcher studied violinists at a music academy and found that the best ones had practised more than the good ones. Malcolm Gladwell picked that up in *Outliers*, gave it a round number and a chapter title, and the ten-thousand-hours rule went into the language: that expertise is a function of accumulated practice, that the number is about ten thousand hours, and that this holds for more or less anything.

The basketball player is the obvious counterexample, and the obvious counterexample is not the argument. One freak does not overturn a rule.

The real argument is duller and much worse for the rule. Ten thousand hours was always an average, and nobody had asked what the spread around it looked like.

Somebody finally did, in chess, where there is an objective rating and you can check. The average time to reach master level came out at about eleven thousand hours, close enough to the famous number that everyone would have nodded and moved on.

The range was three thousand hours to twenty-three thousand.

One man got there in three thousand. Another needed twenty-three. And several who had started as small children put in more than twenty-five thousand and never made master at all.

David Epstein's line about this is the best sentence in the chapter: *the 7,000-to-40,000-hours rule just doesn't have the same ring to it.*

An average with a spread that wide is not a rule. It is a summary statistic wearing a rule's clothing.

Why it made sense to say at the time

The number had escaped into the language the way research findings almost never do. Corporate training decks. Parenting books. A photographer with no golf background quit his job to spend six hours a day for six years trying to reach the professional tour on hours alone.

The underlying research was real, and it was never the problem. What got built on top of it was.

Anders Ericsson, who ran the violin study, never used the phrase "ten thousand hours rule" in his life. He said so in print, pointed at Gladwell's chapter title as the source of it, and used the word *misconstrued*.

And there is one moment in the chapter that shows how completely the idea had won. The Swede, walking the reporter to his train, stops at a bookshop window, points at the Swedish edition of *Outliers*, and says: read this, it is all about your ten thousand hours.

The man about to be used as the world's best counterexample is recommending the book.

Does it hold up

Mostly, and not in the way the book is usually given credit for.

Large pooled analyses since then have found that practice matters enormously in some places and barely at all in others. About a quarter of the difference between people in games. Under one percent in the professions.

So the strong version is dead, and he is right about that. But notice the shape of it. This is not a finding that practice does not matter. It is a finding that how much it matters depends on what you are practising, which is a claim about domains, not about talent.

Then the argument got strange, and this is the part almost nobody outside the field knows.

Ericsson objected that those pooled studies were counting any old hours rather than deliberate practice, and that objection was not unreasonable, because the two are genuinely different things. Deliberate practice in his sense is not playing the piano for an hour. It is working, under a teacher, on the specific thing you are currently bad at, with immediate feedback, repeatedly, and it is unpleasant. Most of what people log as practice is not that. So he re-ran the analysis with a strict definition of what counted.

His filter kept fourteen studies out of eighty-eight. Of those fourteen, the number that came from the professions was zero. The professions being precisely where the other side had found practice explaining almost nothing.

The two camps were computing their headline numbers over different domains, and then argued about them for a decade as though they were rival estimates of the same quantity.

Two smaller things. His best statistic and his largest concession both rest on that one chess study, which has never been replicated in either direction, so if it goes they go together. And Angela Duckworth, whose work he spent a chapter of his next book attacking, now co-authors work making his argument in general form. Her idea was grit: that sustained passion and perseverance for a single long-term goal predicts success better than talent does. His objection was that her most famous results come from places like West Point, where everybody has already been

selected for perseverance, so the finding cannot travel as far as it was made to travel.

Nobody conceded. The argument was not settled. It was absorbed.

So, the verdict from the software: the claim holds, the reasoning holds, and the credit is misallocated. What won was the narrow, correct version, that practice explains a different amount in every domain and that the spread between people is enormous everywhere. What he is popularly credited with is the big version, that talent beats practice. He never said that. Being careful not to say it is the reason he is still standing.

And the basketball player never improved. Not by a centimetre, in all the years since. Which should be uncomfortable whichever side of this you had already picked.

WHEN DOES EXPERIENCE ACTUALLY MAKE YOU BETTER?

There was a physician in New York, some decades ago, celebrated for his diagnostic skill. His speciality was typhoid fever, and his method was to feel around the patient's tongue with his hands.

He was extraordinarily good at it. Again and again he identified typhoid before the patient had shown a single symptom, and again and again he turned out to be right.

He was, as another doctor later observed, a more productive carrier of typhoid than Typhoid Mary, and he was doing it with his hands.

Every diagnosis confirmed his skill. Every confirmation deepened his confidence. He had a tight, fast, high-volume feedback loop, and it taught him precisely the wrong thing, and it would have gone on teaching him the wrong thing forever.

What the book is arguing

Range is built on a distinction that it borrows rather than invents, and everything else in the book rests on it.

Some learning environments are **kind**. Patterns repeat. The rules are stable. Feedback arrives quickly and accurately. Hit a golf ball and you can see immediately whether it sliced, hooked or flew straight, and the next shot presents the same problem again. Chess is kind. So is firefighting in ordinary house fires. So is anaesthesiology.

In a kind environment, experience does what everybody assumes experience does. You do the thing, you find out how it went, you adjust, and thousands of repetitions later you have genuine expert intuition, which feels like magic and is actually a very large library of patterns.

Some environments are **wicked**. The rules are unclear or they change. Patterns may not repeat, or the repetitions may be misleading. Feedback is delayed, or noisy, or absent, or actively wrong.

In a wicked environment, the same process that builds expertise in a kind one builds confidence without skill. The typhoid doctor is the extreme case, but you do not

need anything that dramatic. You just need feedback that is slow or ambiguous enough that you can practise the wrong thing for years without discovering it.

The claim, then, is not that experience is overrated. It is that whether experience produces expertise **depends entirely on the domain**, and that we routinely apply intuitions formed in kind environments to wicked ones without noticing we have changed environments.

Why it made sense to say at the time

Because it comes from an unusually credible source: a peace treaty between two people who had spent their careers on opposite sides.

Daniel Kahneman had spent decades documenting the ways expert intuition fails, and Gary Klein had spent decades documenting the ways expert intuition works, studying firefighters and nurses making fast, correct decisions under pressure that they could not explain afterwards.

In 2009 they co-wrote a paper attempting to identify exactly where they disagreed, and largely failed to find a disagreement. What they found instead was that they had been studying different domains. Klein's firefighters worked in an environment with repeating patterns and rapid feedback. Kahneman's stock pickers and political forecasters did not. Both were right about the domain each had looked at.

That is the finding *Range* is built on, and building a book on it was a reasonable thing to do. It is one of the most-cited papers in judgment research, it resolved a genuine long-running argument, and it produces immediately usable advice: before trusting anybody's experience, including your own, ask what kind of environment produced it.

Does it hold up

The framework is uncontested. What has not happened, in seventeen years, is anybody turning it into a measurement.

The distinction is canonical and qualitative. The Kahneman-Klein settlement stands and is heavily cited, and the kind-versus-wicked formulation has been refined and restated by the researcher who coined it.¹ It gets borrowed constantly, in avalanche forecasting, driver behaviour, medical education, management training.

The borrowing is always qualitative. Nobody has built an index that scores a domain for kindness. There is no measurement of how much feedback delay is too much, or of where a given profession sits, or of the point at which experience stops adding value and starts adding confidence.

This is worth naming plainly. The single most-cited peace treaty in judgment research produced a distinction that nobody has operationalised in seventeen years, and a bestselling book built its architecture on it.

That is not a scandal. It is how frameworks and popular books actually interact, and a qualitative distinction can be entirely correct and enormously useful without ever being reduced to a number. But it means the kind/wicked axis in *Range* is a lens rather than a finding, and it should be read as one.

There is one quantitative expression of it, and the book under-uses its own best evidence.

The meta-analyses of deliberate practice, the ones usually cited as demolishing the ten-thousand-hours rule, break their results down by domain. Practice explained about twenty-six percent of the differences between people in games, about twenty-one percent in music, four percent in education, and **under one percent in the professions.**

That is a single measurement instrument applied across kind and wicked domains, and it produces exactly the pattern the framework predicts. It is the closest thing anybody has to a number for this, and it deserves more prominence than it gets.

And here is where the argument doubles back on itself in an interesting way.

Twenty-six percent of the variance in a single predictor is not a small effect. In psychology it is a very large one. So every number normally cited as evidence *against* deliberate practice is simultaneously the strongest available evidence *for* the claim that in kind domains, narrow repetitive practice produces real, reliable expertise.

The book's concession to its opponents is better supported than its headline.

The chess concession and the headline statistic come from the same study.

Range concedes that early technical specialisation is genuinely necessary in chess, and calls it critical. The evidence is the same questionnaire administered to a hundred and four Argentinian chess players that produced the three-thousand-to-twenty-three-thousand-hours range.

He read it correctly both times. But it means his most-quoted anti-specialisation statistic and his most-quoted pro-specialisation concession rest on one unreplicated study, and if it falls they fall together. Also, the word *critical* is stronger than the source supports: what the study found was a correlation between skill and starting age, and later work complicated even that by showing cognitive ability independently predicts chess skill, which offers a second route to the same correlation.

The chunking story survived, with one correction the book does not carry.

The account of expert memory as a domain-specific library, which collapses when you scramble the patterns, has held up. But the famous random-chessboard reversal is not total. When researchers re-ran those experiments properly, masters retained a small but real advantage even on random boards, which pure chunking cannot explain. The theory was extended to account for it, with a second mechanism of larger retrieval structures.²

So the correct version is that the expert advantage *mostly* disappears when you remove the familiar structure, and the residue is why a second mechanism was needed. *Range* reports the 1973 version, in which the advantage disappears entirely. The residue is small, real, and inconvenient for the strongest form of the specificity claim.

And one claim in the chapter has never been checked at all.

Range argues that experts in kind domains trade flexibility for narrow skill, becoming worse at adapting when the rules change. This is called cognitive entrenchment, and it is intuitively compelling.

It comes from a theory paper, in a journal that publishes theory rather than data. It has accumulated hundreds of citations and, as far as I can establish, no direct empirical test of the prediction as framed.

This is *never revisited*, which is not the same as *not replicated* and is a very long way from *refuted*. The idea may well be true. Nobody has checked, and the honest thing is to say so rather than to let the citation count stand in for evidence.

Finally, the strangest thing about this framework is what its author did with it next.

The kind-versus-wicked distinction is load-bearing for *Range*. It is what licenses the whole argument: breadth pays where feedback is poor and patterns do not repeat, narrow practice pays where they do.

In *Inside the Box*, published in 2026, the word "wicked" does not appear once. The researcher who coined the distinction appears only in a list of people who have died since the previous book. And the framework that occupies the same structural position is a different one entirely, borrowed from Herbert Simon: the difference between well-structured and ill-structured problems.

That would be unremarkable, except that the replacement is deployed with an explicit claim of *generality*. The new book states that constraint-based problem-solving applies to every challenge, and that creativity is simply normal problem-solving pointed at a hard question. The 2026 book contains no domain-partitioning claim at all.

So one book says the rules differ by domain and that this is the master fact. The next book says one mechanism runs everywhere. Neither acknowledges the other.

I am not going to make that a verdict on him, because it is not one. Writers move on, and a book about constraints has no obligation to relitigate a book about breadth. But it does leave a real question about which position to take to the evidence, and the answer is that the evidence supports the first one better than the second. Domain-dependence is the settled position in judgment research. Domain-generality is a claim the 2026 book asserts rather than demonstrates, and we will come to it.

What to take from it

Before you trust anybody's experience, including your own, ask what produced it. Did the situation repeat? Did feedback arrive quickly, and was it accurate? If yes, their intuition is probably worth something and possibly worth a great deal. If no,

their confidence has been built by the same process that would have built it if they had been wrong the whole time.

That is genuinely useful, it is well grounded, and nobody has ever measured it.

WHAT MAKES LEARNING STICK?

Here is an experiment you can run on yourself, and the reason it is worth describing in detail is that it does not work.

Volunteers were taught to recognise the styles of a set of painters. Half studied one artist at a time: six Braques, then six Seurats, then six others. The other half studied the same paintings shuffled together, one Braque, then a Seurat, then something else, in no useful order.

Studying one artist at a time feels enormously better. The pattern comes into focus. By the fourth Seurat you can see what a Seurat is. The shuffled condition feels like being repeatedly interrupted.

Then everybody was shown paintings they had never seen and asked who painted them. The shuffled group was substantially better at it.

That is not the interesting part. Here is the interesting part.

They were shown their own results. They saw, in their own scores, that the shuffled method had worked better for them personally.

And they still said the one-artist-at-a-time method was more effective.

Not most of them, not by a small margin. The overwhelming majority reported that the method that had just measurably failed them felt like the better one, after being shown that it was not.

What the book is arguing

The umbrella term is **desirable difficulties**, and it names a genuinely strange fact about human learning: several of the conditions that make learning feel worse in the moment make it stick better afterwards.

Shuffling topics rather than blocking them. Spacing study out rather than massing it. Testing yourself rather than rereading. Being made to attempt an answer before you have been taught the method, and getting it wrong.

All of these depress your performance while you are learning. All of them improve what you retain later. And the book's point is that this is not a curiosity but a trap, because **the thing being depressed is the only thing you can feel**.

Performance and learning are different variables and they can be pushed in opposite directions. Performance is what you can do right now, with these cues, in this room. Learning is a durable change in what you can do later, elsewhere, when nobody has just reminded you.

From that follows the book's sharpest claim about teaching. The default failure mode of a good teacher is helping. A student stuck on a hard problem is in the exact state where the durable learning happens, and the humane instinct is to give a hint. The hint converts the problem from something the student has to figure out into a procedure the student can execute, and the procedure is what they retain.

The book supports this with a video study of eighth-grade mathematics lessons across seven countries, in which trained analysts coded what happened when teachers set problems that required making connections between ideas rather than following a method. In the United States, when such a problem was set, essentially none of them stayed that way. They were converted into procedures, almost always by a teacher trying to help.

And then the claim that follows from all of it: **early academic head starts fade**. Children given an academic advantage before school tend to lose it, because what was taught was a closed skill that everybody else acquires anyway on the ordinary schedule.

Why it made sense to say at the time

Because this is one of the best-established bodies of work in all of psychology, and almost none of it had reached the people who make decisions about teaching.

The core findings, that spacing beats massing and that testing yourself beats rereading, have been replicated for decades and survive in real classrooms with ordinary teachers, not only in laboratories with volunteers. This is not a fragile literature.

And the book was making a point that is very hard to make convincingly, because everybody's own experience contradicts it. You cannot feel your own learning. You can only feel your current fluency, and fluency is exactly what the good methods reduce.

Does it hold up

Mostly yes, with one carve-out that is larger than the book indicates and one explanation that is one candidate among several.

The illusion is settled, and the demonstration is better than any effect size. The painting experiment is the canonical case and it is the strongest teaching artefact in this literature, because the illusion and its refutation occur inside one person in one sitting. They take the test. They see their results. The feeling does not move.

That tells you the illusion is not ignorance. It is not a belief that can be corrected by evidence, because it is not a belief. It is a perception. Blocked practice genuinely *is* easier, easier retrieval genuinely *does* feel like knowing, and the learner's sense of their own fluency is reporting accurately. Fluency is simply the wrong quantity.

That has consequences which escalate. You cannot trust your own sense of progress. You cannot trust your judgement of your teacher either, because the same fluency cue drives both. And an institution cannot trust the aggregate of those judgements, which is why student evaluations turn out to be a poor guide to how much students learned.

Durability is vindicated. Transfer is contested. And the carve-out is real.

Interleaving does not work everywhere, and the pattern of where it works is informative. In a meta-analysis of the technique, it helped a great deal for learning to categorise paintings, helped modestly for mathematics, and **actively hurt for learning vocabulary**.¹

That is not a failure of the finding. It is the shape of the finding. Interleaving helps when the task is to tell similar things apart, because shuffling forces you to notice the distinctions. It hurts when the task is to associate one thing with one other thing, because then the interruptions are just interruptions.

The technique is not a virtue. It is a tool with a shape, and the book's version is closer to a virtue than the evidence supports.

The teaching finding holds and has been sharpened.

The claim that instructors who most improve current-course performance can harm later performance comes from an unusually good study design: students randomly

assigned to sections, a common syllabus, common exams, centralised grading so instructors cannot inflate their own students' marks, and a mandatory follow-on course so nobody selects out.²

The finding that has survived everything, and which is the one that matters, is that **what students rate is not what students learn**. That is now established three separate ways: by the randomised design, by meta-analysis of evaluations against learning, and by large-scale work finding instructor value-added only weakly related to student ratings.

The specific harm direction, that the best-rated teachers actively damage later performance, is the more striking claim and the less replicated one. It rests on two institutions and should be read as such.

One correction on the mathematics video study, which matters for tone rather than substance. The report's finding is that "less than one percent" of American making-connections problems were taught as making-connections problems. The book renders this as a grand total of zero. It is a footnoted near-floor estimate, not a clean zero, and the finding does not need the rounding.

And on fadeout, the book's explanation is one of at least three.

That fadeout happens is not in dispute, and the field's own strongest statement forecloses the easy dismissal: it is a substantive phenomenon rather than a measurement artefact.³

But *why* it happens is genuinely open. The book's answer is that early instruction teaches closed skills everybody acquires anyway, so the untreated children simply catch up. That is one hypothesis. Another is that the early gains were real but the later environment did not sustain them. A third is that the intervention opened a door rather than building a skill. These make different predictions about what to build next, and the argument is unresolved.

Two things should sit alongside it.

The first is a finding that states the book's thesis better than the book does. Across eighty-six randomised trials with more than fifty thousand participants, **bigger initial effects faded faster**.⁴ That is a quantitative regularity from the fadeout literature itself, and it is a sharper sentence than the one in the book.

The second is uncomfortable and belongs in any honest account. Tennessee ran a randomised statewide pre-kindergarten programme and followed the children. By sixth grade the treated children were performing *worse* than the control group.

Convergence and harm look identical if you only measure the gap closing. That is a reason to be careful with the closed-skills explanation, because it predicts catching up, and catching up is not what happened there.

What to take from it

If it feels easy, you are probably not learning much. If it feels like falling behind, you might be.

Shuffle related material when the task is telling similar things apart, and do not bother when the task is memorising a pairing. Test yourself rather than rereading. Space things out. Sit in the discomfort of not knowing before you are told, because the not-knowing is where the learning is happening.

And do not trust your own sense of how well any of this is going, because that sense is measuring the wrong thing, and being shown the evidence will not change how it feels.

IS GRIT A THING?

Every summer, the incoming class at West Point goes through seven weeks of basic training that the cadets call Beast Barracks. It is designed to be brutal, and a certain number of them quit.

The military had spent decades trying to predict who. They had a composite score built from academic results, physical fitness tests and demonstrated leadership, and it did not do the job.

Angela Duckworth gave the arriving cadets a short questionnaire instead. Twelve items, agree or disagree, about whether you finish what you begin, whether setbacks discourage you, whether you stay interested in the same goals for years.

The score from that questionnaire predicted who would survive Beast Barracks better than the Army's own composite did.

That is a genuinely striking result, it launched a research programme, a bestseller and a TED talk with tens of millions of views, and by the middle of the last decade grit was in school curricula and corporate training decks worldwide.

What the book is arguing

Range does not say grit is fake. It says the famous results come from a specific kind of place, and that this limits what they can mean.

The argument is one statistical idea, and it is the most useful idea in this entire book, so it is worth building properly.

Suppose you want to know whether height matters in basketball. You study NBA players. You measure everybody's height and everybody's scoring, and you find that height barely predicts scoring at all. Among NBA players, some of the best scorers are the shorter ones.

You would be committing an error, and the error has a name: **range restriction**.

The NBA has already filtered for height before you arrived. Everyone in your sample is tall. You have removed almost all of the variation in the thing you are measuring, and a statistic that measures how two things vary together cannot see a variable

that no longer varies. So the correlation collapses toward zero, and it collapses *because of your sample*, not because of anything about basketball.

Read carelessly, your result says height does not matter. Read correctly, it says: among people who are all already very tall, additional height does not distinguish them much. Those are completely different statements, and only one of them is true.

Now apply that to West Point.

Nobody arrives at West Point by accident. They have completed a long, deliberately arduous application process. They have already demonstrated years of sustained commitment to a difficult goal, because that is what the admissions process is designed to select for.

They have, in other words, already been filtered for grit.

So the book's question is: what does it mean that a grit score predicts something inside a population selected for grit? And what would it mean if you tried to generalise from that population to everybody else?

Why it made sense to say at the time

Because by 2019 grit had escaped its evidence entirely.

The original finding was about persistence through a short, defined, externally-set ordeal. What it had become was a general theory of achievement, taught to children, used to explain who succeeds in life, and increasingly used to explain who does not.

That last part matters. A trait-based explanation of success is also a trait-based explanation of failure, and when it is applied to people who did not have the option of persisting, it stops being psychology and becomes a verdict.

Somebody pointing out that the famous results all came from unusual populations was doing useful work.

Does it hold up

The methodological critique is vindicated, and the most interesting support for it now comes from inside the grit programme itself.

On the critique. The external-validity objection is the mainstream reading. And Duckworth has since co-authored work arguing that psychology systematically overstates the importance of single variables, and listing range restriction among the structural reasons that life outcomes resist prediction.

That is convergence, not capitulation. She is making a general methodological argument, not retracting anything. But the substance of the objection is now being made by the person whose findings prompted it, in her own papers.

There is also a detail sitting inside the grit programme's own flagship paper that is the single best illustration of range restriction I know. In a study of over eleven thousand West Point cadets, cognitive ability came out **negatively** correlated with both physical ability and with grit.

In the general population those correlations are not negative. They are negative among cadets for a simple reason: admission is a composite. A cadet admitted with lower test scores must have compensated somewhere else. So within the admitted group, being strong on one dimension predicts being weaker on another, purely as an artefact of how the group was assembled.

That is the signature to learn. **The most diagnostic sign of a selected sample is not a small correlation. It is a negative correlation between two things that ought to be positively related.** When you see one, ask how the sample was assembled.

On the size of the effect, the field's answer is clear and the book's boundary is not the field's boundary.

Grit does predict real outcomes, at modest magnitude. The largest meta-analysis found that the perseverance component correlates with academic performance at about 0.26, and the consistency-of-interest component at about 0.10.¹

Two things follow. The first is that the grit scale is really two things wearing one name, and only one of them is doing work. The perseverance half predicts modestly. The consistency half, which is the part about pursuing the same goal for years and is the more distinctive half of the idea, barely predicts anything.

The second is more damaging. Grit overlaps very heavily with conscientiousness, a personality trait that has been measured for decades. Once you account for

conscientiousness, grit adds little. The field's blunt version is that grit is largely a rebranding.

Now the part where the book and its critics agree on the conclusion by routes that contradict each other.

Range says grit is genuinely predictive where goals are given, measurable and short-horizon, and does not generalise beyond that. That is a claim about **goal horizon** as the boundary condition.

The meta-analytic literature says something different. A synthesis covering more than two hundred and eighty thousand participants tested the proposed moderators and found **no significant moderators of the grit-achievement link at all.**² The boundary is not horizon. The boundary is *facet*: perseverance works broadly at a modest level, consistency works nowhere much, and the whole construct is mostly conscientiousness.

So he reached the right destination, which is that grit does not license the general conclusions drawn from it, by a route the evidence does not support. His fence is in roughly the right place and it is made of the wrong material.

One thing is out of bounds here, and I want to say why.

Range spent a chapter treating Duckworth's work as the foil. In the afterword to the paperback, he withdrew the quarrel: he cites her endorsing early sampling before specialisation, and treats her as an ally rather than an opponent.

Under this book's rules, a position its author has withdrawn does not get audited. He did that work himself and it is not mine to redo. What survives, and what is audited above, is the methodological critique, which he never withdrew and which turned out to be right.

And the general point about traits holds.

The book argues that grit-like traits are situation-specific: the useful question is not whether someone is gritty but when. That framing is uncontroversial as personality theory and has been for decades.

In the grit literature specifically there is supporting evidence, in that domain-specific grit measures outperform the general scale. But those studies are in the hundreds

of participants, mostly from one applied subfield, and should be read as suggestive rather than settled.

What to take from it

Grit is real, small, and mostly conscientiousness under a better name. The half of it that predicts things is the ordinary half about working hard; the distinctive half about pursuing one goal for years is close to inert.

And the larger lesson is the statistical one, which is worth more than the verdict. Whenever a finding comes from an elite institution, a professional league, a selective programme or a group of survivors, the sample has already been filtered on something. Correlations inside that group can be small, absent, or backwards relative to the world outside it.

That is not a reason to dismiss such findings, and I want to be careful here, because the tool cuts too well. Knowing a correlation is distorted by selection tells you the number is not the population number. It does not tell you what the population number is, and treating "your sample is selected" as an argument that wins is a rhetorical move rather than a methodological one.

Use it to ask a better question, not to dismiss an answer you dislike.

CAN ANYONE FORECAST ANYTHING?

In 2011 an American intelligence agency ran a competition to find out whether anybody could predict world events, and invited teams to enter.

One team recruited volunteers off the internet. Retired programmers, a pharmacist, a filmmaker, people who did this in the evenings for the interest of it. They were asked questions with checkable answers on a fixed date: will this leader still be in power, will this currency hold, will these talks produce an agreement.

They were competing, among others, against professional intelligence analysts who had access to classified material the volunteers could not see.

The volunteers won. Not narrowly, and not once.

What the book is arguing

The finding underneath that tournament is older and more uncomfortable, and it comes from a twenty-year study in which 284 people who made their living commenting on political and economic trends were asked to put probabilities on specific outcomes. Eighty-two thousand forecasts.

The results: the average expert was a poor forecaster. Fifteen percent of the events they declared impossible happened anyway. More than a quarter of the things they called sure things did not occur.

And credentials did not help. Specialty did not help, experience did not help, degrees did not help, and access to classified information did not help.

What *did* predict accuracy was a style of thinking, and the book uses the old distinction between hedgehogs and foxes. The hedgehog knows one big thing and drives every question through it. The fox knows many small things, borrows from several incompatible frameworks, holds views loosely, and updates.

Foxes forecast better. And worse for the hedgehogs, the study found they got *less* accurate as they accumulated credentials and experience, and were at their worst inside their own speciality, where the single big idea had the most grip.

The book's second claim is the more useful one: **this is trainable**. Not a personality you either have or lack. A set of habits that can be taught.

And its third is about the moment when a familiar tool should be dropped and is not. The book collects cases of experienced professionals under pressure reverting to the method they know, including firefighters who would not put down their tools while running from a fire, and organisations that demanded a familiar kind of analysis at exactly the moment the familiar kind could not answer the question.

Why it made sense to say at the time

Because the tournament result is genuinely startling and it had barely reached the public.

And because the alternative account of forecasting, the one everybody carries around, is that prediction is a matter of knowing more. If that were true, the analysts with classified access would have won. They did not, and the reason they did not is worth a book.

Does it hold up

Very well on the parts that were measured, and much less well on the parts that were decorated.

The tournament findings hold, and the corroboration is unusually strong. The original twenty-year study is one research programme's archive and has never been replicated as a dataset. But its central claim was independently confirmed by a separate, adversarial, government-run tournament a decade later, in which the recruited amateurs beat the credentialed professionals. Two different research groups, different decades, different methods, same answer.

The training result is the single strongest position in this part of the book, and it deserves the number.

Inside that tournament, forecasters were randomly assigned to receive a short training module in probabilistic reasoning: think in reference classes, break questions into parts, distinguish what you know from what you assume, update in small increments.

The training lasted **less than an hour**. It improved forecasting accuracy by **six to eleven percent**, against a randomised control, consistently, for four consecutive years.¹ The authors note these are probably lower-bound estimates.

If you want one piece of evidence that good thinking is a set of habits rather than a gift, that is it, and nothing else in this literature comes close.

The fox and hedgehog vocabulary, though, has quietly gone.

Everything empirical about the position held: breadth helps, integrating incompatible frameworks helps, willingness to update helps, and teams of good forecasters beat crowds, prediction markets and professional analysts. That is all confirmed.

What did not survive is the label. The successor research does not measure whether somebody is a fox. It measures the things the fox description was gesturing at: active open-mindedness, frequency and size of belief updating, granularity of probability estimates. Those are measurable and the metaphor is not.

I should be careful here, because this is my inference rather than a finding. Nobody published a paper arguing the distinction is invalid. It simply stopped appearing in the predictor sets, which is how constructs usually die, and it is a weaker kind of evidence than a refutation.

One of the supports has failed outright, and it is worth naming because it is a good lesson.

The claim that credentialed hedgehogs get *worse* with more information leans partly on a famous experiment. People were given a set of numbers about the effectiveness of a skin cream and asked to interpret them, and they did fine. The same numbers were then relabelled as being about gun control, and suddenly the more numerate participants were the ones who got it wrong more often, in the direction of their politics. The interpretation was that intelligence gets recruited to defend commitments.

That experiment was re-run in 2023 with a nationally representative American sample, and it **did not replicate**.² It half-failed in Europe too.

The tournament finding stands on its own. The skin cream study should come out. And a general claim about clever people and motivated reasoning should now be made, if at all, with the acknowledgement that its most-cited demonstration failed to reproduce.

That is a better paragraph than the one it replaces, and the book would be stronger for it.

The Challenger argument is the position most likely to be wrong in three books, and I think it is wrong.

Range uses the space shuttle *Challenger* as its central case for over-reliance on quantitative analysis. The night before the launch, engineers argued the cold weather was dangerous. Management asked for data. The engineers had charts of O-ring damage from previous flights that did not show a clean temperature relationship, and could not produce the number in the form demanded.

The book's reading is that this was a failure of *insisting on* quantification: the relevant data could not have settled the question, and the engineers' qualitative judgement was the better guide.

The statistical literature does not support that. The reference analysis, published three years after the accident, used all twenty-three pre-accident launches, including the seventeen with no O-ring damage that had been excluded from the charts on the night. It found strong statistical evidence of a temperature effect, and estimated at least a thirteen percent probability of catastrophic O-ring failure at the launch temperature.³

The data *could* have settled it. What defeated the engineers was that they were looking only at the flights where something had gone wrong.

That is a selection error, and it is the same error this book keeps running into: if you look only at cases where the outcome occurred, you cannot see the relationship. Excluding the undamaged flights *was* the mistake, and including them *was* the fix.

So the defensible version of the chapter, which keeps everything valuable in it and loses the error, is this: NASA demanded a number, would not accept an argument, and the engineers who were right could not produce the number in the form demanded on the timescale available. That is a real and damning story about organisations. It is not a story about the limits of quantitative analysis, because the analysis that would have settled it existed and simply had not been done yet.

On the organisational culture material, one claim needs correction.

Range reports a study of 334 institutions finding that cultural congruence had no influence on any measure of organisational success whatsoever.

The paper is identifiable and its own abstract says something different: that the findings provide support for both lines of inquiry, and that the effect he describes as absent is present, though weaker than the other effect they tested. I could not obtain the full text, so I cannot rule out that one specific measure was null while the headline reported support.

What I can say is that "no influence whatsoever" is stronger than the source will carry, in a chapter arguing that people overstate what their evidence shows.

And one finding that holds, with a clause the book should have added.

Range reports that expeditions from countries with more hierarchical cultures got more climbers to Himalayan summits and also got more of them killed, and that this did not hold for solo climbers, only teams.

That is faithfully reported from a study of more than thirty thousand climbers on over five thousand expeditions, and the solo-climber detail is the study's own and is what makes it about group dynamics rather than about recklessness.⁴

The clause worth adding is that the hierarchy measure is *national*, not measured on the expeditions themselves. It is a country-level average applied to teams from that country. That is the difference between a striking finding and an overclaim, and it costs nothing to say.

What to take from it

Expertise and forecasting are different skills, and the second one is rarer and more teachable than anybody expects. Credentials do not supply it. An hour of the right instruction measurably does.

The habits are unglamorous: break the question up, ask how often this sort of thing happens in general before asking about this particular case, put a number on it so you can be wrong in public, and move the number in small steps when evidence arrives.

And the deeper thing, which runs under this chapter and the Challenger case and half of this book: **when you look only at the cases where something went wrong,**

you cannot see what causes it. The engineers had that problem, the grit researchers had it, the talent scouts had it. It is the most common way that careful people reach confident wrong answers, and it is almost invisible from the inside.

SHOULD YOU KEEP YOUR OPTIONS OPEN?

An economist put up a website for people who could not decide something major. Should I quit my job. Should I end this relationship. Should I move.

Twenty thousand people came. Two thousand of them had a decision big enough to qualify, and the site flipped a coin for them. Heads, make the change. Tails, stay.

Nobody was obliged to obey the coin. But enough people did that six months later there was something to measure, and what it measured was that the ones the coin had pushed toward change reported being substantially happier than the ones it told to stay.

Whatever we are collectively doing when we agonise over these decisions, we appear to be biased toward staying, and the bias appears to cost us.

What the book is arguing

The idea underneath *Range*'s advice about work is **match quality**: how well a person's abilities and interests fit what they actually do.

The argument runs like this. Specialising early buys you a head start in accumulated skill. Sampling widely first buys you information about yourself. Both are real, and they trade against each other.

The book's claim is that the information is usually worth more than the head start, for a reason that is easy to state and hard to accept: **you cannot work out your match quality by thinking about it**. Not by introspection, not by career assessment questionnaires, not by imagining yourself in the role. You find out by doing things and noticing what happens, which necessarily takes time and necessarily means some of that time is spent on things you will abandon.

From that follow two conclusions that run against everything anybody tells a twenty-two-year-old.

The first is that **quitting is often rational**. If the point of trying something is to gather information about fit, then discovering a poor fit is a *success* of the process, not a failure of character. Persisting past that point is not grit. It is ignoring your own data.

The second is stranger and better. Personality and preferences change substantially through adulthood, most of all between eighteen and the late twenties. We systematically underestimate how much we will change; people of every age reliably report that they have changed a great deal up to now and expect to change little from here, which cannot be true of everybody.

So specialising early is not just a bet on an uncertain future. It is a prediction about a person who does not exist yet, made by somebody who has systematically underestimated how different that person will be.

Why it made sense to say at the time

Because it ran against an enormous amount of institutional pressure, and because there was a genuinely good natural experiment behind it.

English and Welsh university students have to choose a narrow specialism before they arrive. Scottish students take a broader course and specialise later. An economist compared the two systems and found what the theory predicts: the early specialisers earned more at first, and the late specialisers caught up, because they switched fields less often afterwards. When the late specialisers did switch, they were switching from a better-informed starting point.¹

That is a clean comparison of two systems in the same country doing the same thing differently, which is about as good as this kind of question gets.

Does it hold up

Unevenly, and the unevenness is instructive, because the parts that hold are the careful ones and the parts that wobble are the quotable ones.

The match-quality mechanism is mainstream. That people do better in work that fits them, and that fit is discovered rather than deduced, is not controversial.

The specific natural experiment has never been replicated. It remains one clean study, correctly reported, in one institutional setting, and everything the book builds on it rests on that one study. That is a mark against the field more than against him, but a reader should know that the empirical floor under this argument is a single paper.

One claim reports the wrong outcome variable.

Range says that narrow, career-focused education produces early advantages that reverse into lower lifetime *earnings*, and that the faster an economy changes, the bigger the long-run advantage of a general education.

The paper it draws on found a reversal in **employment**, not earnings.² Vocationally trained workers are more likely to be employed when young and less likely to be employed later in life. That is a real and important finding and it is not the one described.

The earnings version is separately supportable from other work, which the book does not cite. And there is a further problem it does not mention: a later paper reproduced the employment reversal while eliminating the explanation that makes it a story about narrowness at all. The reversal is real; the reason for it is not settled.

The most quotable line in this territory is the least supported.

"Switchers are winners" is the kind of sentence that ends up on a slide. The dramatic version of it rests on the coin-flip study, and that study has a structural problem the chapter does not report: the participants were self-selected people who went to a website about a decision they were already agonising over, and then chose whether to obey a coin. That is a striking piece of work and it is not a random sample of anything.

The narrow version, that switching is often beneficial and that people under-switch, is defensible. The slogan is not a finding.

On personality change, the two halves come apart.

That personality changes substantially through adulthood is settled, and a large synthesis of longitudinal data has if anything strengthened it: stability of individual differences does not keep increasing after twenty-five, as the earlier consensus held.³ He could have said more than he did.

The end-of-history illusion, the claim that we systematically underestimate our own future change, is a different matter. It comes from a famous study, it has been extended to children and moderated by culture, and it has never been confirmed by the one design that could test it, which is following the same people forward and comparing what they predicted against what happened.

Not refuted. Not replicated in the form the argument needs.

And the revision to the personality-change picture cuts both ways for him. If stability does not keep rising after twenty-five, then specialising at eighteen is a bet on an unformed person, which is his point. But so is specialising at thirty. The implied promise that you should wait and then you will know does not survive its own premise, because there is no age at which the changing stops.

One thing here is exactly right and gets better with time. *Range* argues that the marshmallow test has been overread, that a four-year-old's ability to wait for a second sweet has been made to carry far more than it can. The larger and more diverse replication found the effect is real but roughly half the original size, and mostly explained by family background and early cognitive ability once you control for them.⁴ Overread, not refuted. He used the word *misconstrued*, which was the right word then and is the right word now.

And then the same author argued the opposite

Here is the thing this chapter cannot avoid.

Range is a case for keeping options open long enough to find out what fits. Its 2026 successor, *Inside the Box*, repeatedly praises closing them.

That book reports research on what it calls sliding versus deciding, and treats drifting into commitments without choosing them as the failure mode. It reports that people prefer reversible decisions and yet end up more satisfied after irreversible ones. It recommends commitment devices, fixed dates, closed doors. It argues for consolidating what you care about into a small set of things. And it contains a personal passage, dated to the period just after *Range* came out, in which the author describes having spent a decade optimising for autonomy, achieving an unusual amount of it, and discovering there is such a thing as too much.

He does not mention the tension. There is no paragraph reconciling the two.

I want to be careful about what to do with that, because the easy move is to call it a contradiction and treat that as the finding, and it is not the finding. Writers move on. A book about constraints is under no obligation to relitigate a book about breadth. And there is a reading on which both are right: sample widely to find the

thing, then commit hard to it once found. He gestures at that once, in a single sentence about a chemist, and never returns to it.

What matters here is not his consistency. It is which position to take to the evidence.

And the evidence, such as it is, does not settle it. The option-value literature supports sampling before commitment. The satisfaction-with-irreversibility literature, which we will get to, turns out to be considerably weaker than his newest book presents it, and the paper he cites for its central claim reports a moderation that undoes it.

So the honest answer to the chapter's question is: keep them open while you are still learning something from having them open, and understand that this is a much narrower claim than either of his books makes.

What to take from it

You cannot think your way to knowing what suits you. That part is solid and it is the useful part.

Quitting something that has taught you it is a poor fit is not weakness, and the sense that it is comes from a story about persistence that the evidence does not support.

You will change more than you expect, and there is no age at which that stops, which means no amount of waiting produces certainty. It only produces better information than you had.

And the person telling you all this has since written a book arguing that too many open options is its own trap. He is probably right about that too, which is inconvenient for anybody who wanted a rule.

DOES BREADTH ACTUALLY PAY?

There are companies whose entire business is taking a problem that has defeated a corporation's own scientists and posting it on the internet for strangers.

The strangers solve a substantial fraction of them. And when researchers looked at who solved what, they found that **the further a solver's own field was from the field the problem came from, the more likely they were to crack it.**¹

That is a strange and appealing result and it is the emotional core of *Range's* argument about work.

Before going further, one thing about it that the book does not report and that you should hold on to. Only problems that have *already defeated the insiders* get posted. That is the selection criterion for the entire dataset. So the solve rate cannot be compared to any insider solve rate, because the insiders' successes never enter the sample. What the study shows is that outsiders can solve problems insiders could not. It cannot show that outsiders are better at solving problems.

What the book is arguing

The claim has several forms, and separating them turns out to be the whole game.

The weakest and most quotable: generalists beat specialists. Breadth wins.

The version the book mostly argues: the modern world is fast-changing and uncertain, and in uncertain domains breadth pays while in stable ones depth does. That is a moderated claim, and it is the book's own fence.

The version with the most interesting evidence: work that combines previously unconnected bodies of knowledge is the work that eventually matters most.

And the version about people: individuals who accumulate experience across many fields, then apply it, out-invent those who went deep in one. The book has a name for the winning type, borrowed from a researcher who studied inventors at a large American manufacturer: the polymath, broad across many technology classes with depth in at least one.

Why it made sense to say at the time

Because the mechanism is real, several of the studies are excellent, and the cultural pressure runs entirely the other way. Every institution a person passes through rewards narrowing. Somebody pointing out that the biggest payoffs come from connections between fields was saying something true and useful.

Does it hold up

This is the chapter where assembling all the evidence in one place changes the verdict, and the change is worth the trouble.

One proposition in this family is superbly well established, and it is not the one the book is famous for.

The claim that combining previously unconnected bodies of knowledge produces work that is **punished early and disproportionately rewarded late** is about as solid as social science gets.

It rests on an analysis of nearly eighteen million scientific papers, and a separate one of a million more.² Highly novel papers have more than triple the probability of ending up in the top one percent of citations on a long time window. They are also published in lower-impact journals and under-cited in the short term. The same pattern shows up on the funding side: novel proposals are less likely to be funded.

And the critical methodological point: **these studies are not selected on success.** They observe entire literatures, including all the papers nobody ever read. That is what makes them different from almost everything else in this territory.

So the well-evidenced claim is: unusual combinations pay off, the systems that evaluate work punish them early, and the payoff arrives on a timescale longer than any grant cycle or citation window.

That is a claim about **combinations**, not about **people**. Hold that distinction, because everything turns on it.

The claim about individual careers is the weak link, and it is weak in four separate places.

First, the sentence in the Introduction that carries the entire book out of sport and into the world of work. It says a raft of studies show that technological inventors increase their creative impact by trading a little depth for breadth as their careers progress, and that there is a nearly identical finding for artistic creators.

I went through the endnotes for that section of the book. Neither claim has one. The surrounding sentences all carry endnotes; these two do not.

And the literature that does exist on career trajectories finds close to the **reverse** sequence: creative people tend to explore first and then narrow, with the narrowing coinciding with their most productive period.³

Second, the polymath study itself. It is real, and it splits its result three ways, and the split is not the one the chapter reports. In the paper's own words: breadth of inventor expertise relates to generating *many* inventions, **but not necessarily technically influential ones**. Depth of expertise is what enables technically influential inventions. Breadth-plus-depth wins on commercial conversion.

So on technical impact, depth wins. That is in the source.

Third, the historical trend that would turn this into a law. The chapter reports that the contribution of specialists to invention peaked around 1985 and has declined since, from an algorithm run over ten million patents.

That analysis is unpublished. It is attributed in the text, by name, to the researcher who ran it, which is honest, and it cannot be checked. The published literature on the same question points the other way: innovators have been *narrowing* over time, as fields grow and the burden of what you must know to reach the frontier increases.

That researcher is also the chapter's protagonist, the exemplar of the winning category, and the source of its historical trend. Four roles.

Fourth, the claims about breadth in creative work rest on a single 2006 study of a single industry, with no replication anywhere, extended by analogy to famous individuals chosen because they succeeded.

And there is a direct counter-result the book does not engage, published the same year.

Researchers used the collapse of the Soviet Union as a natural experiment. When the archives opened, a large body of Soviet theoretical mathematics suddenly became available to Western mathematicians, and the pace of change in affected subfields jumped.

What happened to generalists and specialists? **Generalists performed best when the pace of change was slow. Specialists gained the advantage when the pace of change increased**, and specialists became the more sought-after collaborators under fast-moving conditions.⁴

That is close to the opposite of the book's framing device, which is that breadth wins precisely in fast-changing uncertain environments.

The result Epstein does cite for his moderator is observational. This one is identified, which means it has a cleaner claim to causation. It appeared in 2019, the same year as *Range*, and does not appear in it.

One more thing, and it is the most useful structural observation in this chapter.

The inverted U shows up three times in the underlying literature: in team structure, in the relationship between slack resources and innovation, and in the relationship between novelty and impact. In each case the answer is not "more is better" but "there is an optimum and both extremes are worse."

It appears zero times in the book.

That is not a small stylistic point. Reporting the curve would have made the argument *better*, because an optimum is a more useful and more defensible claim than a direction, and it is what the evidence actually shows. The rhetoric is monotone where the evidence is curved.

The consolidated answer

The evidence supports **breadth of knowledge inputs into a piece of work** far more strongly than it supports **breadth of a person's career**.

Recombination across distant fields is well measured, cross-domain, replicated, and observed on unselected samples of entire literatures. It is one of the better-established findings in the science of science.

Individual breadth versus depth is contested in the field's own words, has one identified natural experiment that favours specialists under fast change, and in this book rests on one company study whose own conclusion gives depth the technical-impact outcome, one comic-book study nobody has replicated, and one unpublished patent analysis by the chapter's own protagonist.

The trade-book version, be a generalist and you will out-innovate the specialists, is not what any of this shows.

The defensible version is narrower and is still worth a book: **the combinations that pay off are the ones nobody has made yet; the systems that evaluate work punish them early; and the reward arrives on a timescale longer than anybody's career incentives.**

That version is fully cited. It is not selected on success. And it is the version *Range* itself states best, in one sentence, deep in a later chapter, where almost nobody quotes it.

DO CONSTRAINTS MAKE YOU BETTER?

Everybody knows that Dmitri Mendeleev discovered the periodic table in a dream.

He did not. The book's newest volume opens by taking that story apart. Mendeleev was writing a textbook, he was behind, and he had a chapter to finish and a limited number of elements he could fit into the structure he had. The organising insight came out of a publishing constraint.

That is a good opening move, and worth noticing for its own sake: a book about how things get made begins by demolishing the most famous origin story in its own subject, on the grounds that the evidence does not support it.

The other story it opens with is a pianist arriving for a late-night concert to find the wrong instrument on stage. The pedals barely worked, the upper register was thin and tinny, the whole thing was too small for the hall. He nearly refused to play. Because the top notes were unusable he stayed in the middle of the keyboard, and because the instrument was quiet he hammered it, and the recording of that night became the best-selling solo piano album ever made.

The defect made the record.

What the book is arguing

The thesis of *Inside the Box* is that limits improve creative work, and that we systematically believe the opposite.

The mechanism offered is that human beings are cognitive misers. Given a completely open problem, we do not explore the space of possibilities; we take the first familiar route, because thinking is expensive and the familiar route is cheap. What a constraint does is block the familiar route, and force exploration of the rest.

From that come several more specific claims. That total freedom paradoxically produces conformity, because everybody defaults to the same obvious solution. That the useful kind of constraint is a *pair*: preclude a familiar answer and promote a specific replacement. That prescribing both the goal and the exact process is the wrong kind of limit and kills the benefit. And that hard problems are what a Nobel laureate called ill-structured, meaning the space of possible approaches is unbounded, so the solver's first job is not to search the space but to shrink it.

There is a second argument running alongside, about science itself. That researchers who collect data first and form the hypothesis afterwards are not testing anything, and that requiring them to state predictions in advance transformed a field: the proportion of large American cardiovascular trials reporting a benefit collapsed after registration became a condition of funding.

Why it made sense to say at the time

Because it is the counterintuitive version of something people half-know, and because the mechanism is easy to verify against your own experience of a blank page.

And because the preregistration story is real and important, and had not reached a general audience.

Does it hold up

This is the newest book, and the honest answer requires changing the question.

Most of it has not been tested. Most of it never will be.

Of the forty-two positions in the first half of the book, nineteen are untested by anybody. Nine are contested. Eleven are unremarkable statements of what the relevant field already thought. Three are vindicated.

None is superseded. And that is the informative number. Nothing in the book has been overtaken, because almost nothing in it has been engaged with. It is two years old.

Which means the audit that matters for a new book is a sourcing audit, not a replication audit. You cannot ask whether a claim replicated when nobody has run it once. What you can ask is: what was in the notes, how many studies, what design, which field, what does that field's own meta-analysis say, and did the author cite it.

On that test the results are uneven in a specific and predictable way. The chapters on history and on research methods do well. The chapters on creativity do less well, and the pattern there is consistent: a chapter thesis supported by five small laboratory studies, with a meta-analysis on the same question left uncited.

Here is the finding that should change how you read any new book of this kind. Recency is not credibility, and in this book recency correlates with weakness.

The best-supported material in it is old. A study of problem-solving fixation from 1942. A study of imagination from 1994. The choice-overload meta-analysis from 2010. The discovery-learning meta-analysis from 2011. The atypical-combinations paper from 2013. The problem-space framework from the 1970s.

The most fragile material is newer: a small creativity study from 2010, an organisational study from 2018, base rates about megaprojects that come from a trade book rather than a paper.

A reader who assumes a 2026 book is built on 2020s science has it exactly backwards. The old material is old because it survived.

Three papers would have made the book stronger. All were published before it. None is cited.

There is a meta-analysis of the book's own central thesis, covering more than a hundred studies, and it supports him. It also concludes that methodological artefacts better explain the variation in existing findings, which is a more interesting result than a simple confirmation.

There is a meta-analysis of a hundred and sixty-four studies of discovery learning, which proves one of his claims outright: unassisted discovery does worse than explicit instruction, while *guided* discovery does better than both. That is his preclude-and-promote structure, demonstrated at scale, seventeen years before he argued for it.

And there is a 2025 *Nature* paper quantifying exactly the effect his chapter on radical novelty describes.

Now the one outright factual error I found in three books.

The book reports that founders trained to treat their business as a scientific experiment pivot more and earn more. The source is a randomised trial of a hundred and sixteen Italian startups, published in 2020.

The same authors ran it again. Four further randomised trials, seven hundred and fifty-nine firms, published in 2024, open access, two years before *Inside the Box* appeared.

Their sentence: **this result did not replicate in the subsequent trials.**¹

What replaced it is that treated founders pivot *once*, rather than not at all or repeatedly. The revenue effect survived but small, about seven thousand euros against a mean of under sixteen thousand with a standard deviation over eighty thousand, which is a distribution where a difference in means is a fragile thing to build a paragraph on. He notes the skew himself, which is more than the sentence required.

And the effect that actually strengthened in the replication is one the book does not mention: treated founders terminated bad ideas earlier.

I want to be precise about what this shows, because the obvious reading is wrong. It does not show the science was bad. It shows the opposite. The people who broke the finding were the people who found it, they went looking deliberately, and they published the part that made them look worse. That is what a healthy field looks like from the inside: not the absence of a broken result, but the presence of the people who broke it.

What it shows about the book is narrower and still a problem. The correction was public, free, two years old, and by the same lead author.

On the preregistration argument, the evidence is real and the language slips.

The collapse in positive results after registration became mandatory is a real finding, and there is a second, independent version of the same shape: comparing journal articles where the study design was reviewed before data collection against ordinary ones, the proportion reporting positive results falls from about ninety-six percent to about forty-four percent.

What the original study says is that registration was **strongly associated** with the trend toward null findings. Association. Its design is an observational time series with fifty-five trials in which the treatment is perfectly confounded with the calendar, and several other things changed at the same time: the standard of care improved, which shrinks any new drug's incremental benefit; the mix of trials

shifted toward dietary supplements, which have a near-uniform null record; trials got larger and better powered; and the easy wins had already been found.

Across a single chapter, the same inference appears three times at three different strengths. The earlier positive results were "**presumably**" false positives, which is correct and correctly qualified. The restriction "**caused**" the shift. And "**undoubtedly**" there are ineffective products on the market because of researcher freedom.

Presumably, caused, undoubtedly. One inference, one chapter, three epistemic strengths, and they get stronger as the chapter goes on.

That is a specimen worth keeping, because it is not carelessness. It is what happens to a careful writer's hedge under narrative pressure over a few hundred words.

Which brings up the pattern this whole book has been running into, for the sixth time.

The hedges hold. On the central thesis and on five other positions in this section, he installs precisely the boundary condition a critic would demand, usually within a page of the claim, often in his own voice rather than attributed to somebody else. One footnote goes further and states his method outright, including an admission that he formed hypotheses after seeing data as a graduate student, that he used a tool to search for contradictory findings, and that no such effort is ever nearly perfect. That is more epistemic self-disclosure than most books of this kind contain anywhere.

Where he breaks, he breaks in the same place every time: the sentence at the end of a section, where the hedged claim is restated as a slogan. Total freedom is the enemy of creativity. The way to instantly make someone more creative is to force them to explore by limiting their options.

The claim in the paragraph is defensible. The claim on the pull-quote is not.

Anybody quoting this book should quote the paragraph and never the line, and that is true of all three of his books.

And one thing belongs here that cuts the other way entirely.

There is a famous study, cited everywhere in education and urban design, in which children in a playground without a fence huddled near their teacher, and children in a fenced playground used the whole space. It is a perfect illustration of this book's thesis. It would have been the best story in it.

The book reports that the study does not exist. And it is right: it cannot be found in any database, and its trail runs into secondary citations that lead nowhere.

He went looking for his own best illustration, could not source it, and printed the absence instead of the anecdote.

That deserves to sit next to the Camuffo error, not instead of it. Both are true, and a portrait with only one of them in it would be a false portrait.

What to take from it

Constraints probably do help creative work, in a way that is real but has an optimum rather than a direction, and the honest state of the evidence is that most of the specific claims in this book have not been tested and several rest on single studies from fields with poor replication records.

The more useful thing is the reading lesson. When a new popular science book arrives, you cannot ask whether its findings replicated. You can ask what is in its notes, whether the studies are large, whether the field has a meta-analysis on the same question, and whether the author engaged with it. That is a check a reader can actually run, and this book mostly passes it in some chapters and mostly does not in others.

And the oldest material in it is the strongest, which is not a coincidence and is the single most portable thing in this chapter.

WHAT SHOULD YOU SETTLE FOR?

There is a famous experiment involving jam.

A supermarket set up a tasting table. Sometimes it offered six varieties, sometimes twenty-four. The larger display attracted more people to stop and taste. The smaller display sold roughly ten times as much jam.

The conclusion everybody took away is that too much choice paralyses us, and it became one of the best-known findings in consumer psychology. It is quoted in books about decision-making, in talks about product design, and in arguments about everything from restaurant menus to pension schemes.

It is also, as a general law, not really there. We will come back to that.

What the book is arguing

Inside the Box closes with its most personal and most prescriptive section, and it argues four things in sequence.

Stop searching and settle. The formal case comes from Herbert Simon, who coined the word *satisficing*: because no real person can evaluate every option, the rational policy is not to optimise but to set a standard, search until something clears it, and stop. Trying to maximise is not merely difficult, it is unavailable, and pretending otherwise costs you the search time.

Prefer commitments you cannot undo. People say they want reversible decisions and reliably end up more satisfied after irreversible ones. Leaving the door open keeps the question alive, and an alive question is a cost you pay daily.

Accept obligation. Rulelessness is corrosive. Being bound to other people, in ways you did not choose and cannot easily exit, is protective rather than limiting.

And constrain your own autonomy deliberately. Synchronise your time with other people rather than optimising your own schedule. Choose a small number of things you care about and stop renegotiating them. There is a personal passage here, dated to just after his previous book came out, in which he describes having spent a decade pursuing autonomy in his work, having achieved an unusual amount of it, and having discovered there is such a thing as too much.

Why it made sense to say at the time

Because the formal argument is unimpeachable and almost nobody acts on it.

Simon's point is not psychology, it is arithmetic. A bounded agent cannot evaluate an unbounded option set, so a stopping rule is not a compromise with rationality, it *is* rationality once you count the cost of searching. That has never needed defending and does not now.

And the surrounding cultural pressure runs hard the other way. Everything about modern life, and modern software in particular, is built to keep the option set open and the search running.

Does it hold up

About half of it, and the half that holds is not the half the prose leans on.

The strongest leg is the one he leans on least.

The claim that obligation and connection are protective is supported by some of the best evidence anywhere in these three books: a synthesis of a hundred and forty-eight studies and more than three hundred thousand participants, finding a substantial survival advantage for people with stronger social relationships, stable across age, sex, baseline health, cause of death and follow-up length.¹

That evidence sits in a footnote. The body text is given to Durkheim, whose nineteenth-century suicide statistics have not been usable as evidence for about a century.

The argument would be stronger, not weaker, if the citations were the other way round. This is the only place in three books where I would tell him to swap a footnote with his main text.

The weakest leg is carrying the most weight.

Satisficing is where the closing argument begins, and it rests on three propositions of wildly different quality.

The formal one is Simon's and needs no defence.

The second is that abundant choice actively harms us, and that is the jam study, and it is the one that has deflated. A meta-analysis of the choice-overload literature

found a mean effect near zero.² A later rescue operation restored an effect through four moderating conditions, and concedes in its own abstract that whether and when large assortments impede choice remains an open question.

Some specific results survive. Adding fund options to a retirement plan really does reduce participation. What does not survive is the general law, which is what everybody took away and what the closing argument needs.

The third is that maximisers, people who try to find the best option rather than a good one, are worse off: less satisfied, more regretful, more prone to comparison. That survives meta-analysis at real scale, six hundred and eighty-three effect sizes across more than forty-seven thousand people.

But it rests on a construct that may not be one thing. There are around eleven competing measures of maximising in print, and the co-author of the original scale has publicly described the resulting literature as befuddling and contradictory.³

When you split the construct, the finding splits with it. Maximising as a *goal*, meaning having high standards, shows no relationship with wellbeing. Maximising as a *strategy*, meaning endless searching among alternatives, is where all the harm lives.

That distinction rescues his practical advice, because his advice targets the searching. But a reader who takes away "do not be a maximiser" has taken away a claim about a construct the field no longer thinks is a single thing, and has possibly concluded they should lower their standards, which is not what the evidence says at all.

And then the hinge, which has a specific and checkable problem.

The claim that irreversible decisions leave people more satisfied is what converts satisficing from a search heuristic into a philosophy of life. If closing the door actually feels *better*, then commitment is not merely efficient, it is pleasant, and the whole closing argument follows.

The 2002 finding behind it is real.

But there is a 2015 paper that tested the effect for moderation, and found that it reverses. Satisficers were more satisfied after an irreversible choice. **Maximisers were more satisfied after a reversible one.**⁴

The effect flips by exactly the individual difference the surrounding chapter is about.

And the book cites that paper. It cites it in the footnote where it takes a small closing detail about extreme maximisers, and leaves out the headline result.

That is the sharpest single finding in this audit, and it matters because the sentence it undercuts is the one licensing the advice the chapter closes on. It is also a different failure from the pull-quote overreach that shows up everywhere else in these books. This is citing a paper for a detail while omitting the finding that would complicate the argument, and it happened four times across the two halves of this book.

The autonomy material is honest and thin, and he says so.

The passage about having had too much autonomy is a sample of one, and he labels it as one with unusual discipline, bounding the claim in both directions. The argument that synchronising with other people matters more than controlling your own schedule rests on a clever national time series that never measured schedule control, so it cannot actually support the comparison. And the closing material about choosing a few values and not renegotiating them is philosophy, sourced to philosophers, hedged with *I think*, which is the correct register and deserves credit rather than audit.

The net

The closing argument is roughly half supported, and the supported half is not the half the prose emphasises.

Obligation and connection: strong, and buried in a note. Satisficing as a search discipline: unimpeachable as theory, weak on the choice-overload premise, unstable on the maximiser construct. Irreversible commitment: undercut by a paper he cites himself. Autonomy: untested and honestly labelled as such.

Which produces an unusual verdict. **A reader could follow this chapter's advice and be doing mostly sensible things for mostly wrong stated reasons.**

Set a standard and stop searching: right, for Simon's reason, not the jam study's. Commit rather than keep the question open: probably right for many people, and specifically wrong for the people most inclined to keep searching, which is exactly the group the chapter is addressed to. Stay bound to other people: right, and better supported than anything else in the book. Do not optimise your schedule into solitude: probably right, and untested.

And one thing worth saying about the shelf

If you own both of his last two books, you will notice that they point in opposite directions.

The 2019 book argues for sampling widely, delaying commitment, and keeping options live long enough to learn what fits. The 2026 book argues for foreclosing options, committing irreversibly, and accepting obligation.

Both can be true. The first is about how to acquire information about yourself; the second is about what to do once you have it. He signals the transition once, in a single sentence in the introduction, asking how one channels a broad toolbox into achievement.

But he never argues it. The switch is presented as a maturation rather than as a boundary condition, and the pull-quotes of the two books contradict each other flatly.

I have no verdict on which is right and this book does not adjudicate between an author and himself. What I can report is the asymmetry in the evidence: the case for sampling rests on a clean natural experiment and a large personality-change literature, and the case for commitment rests on a deflated choice literature and a paper whose headline he did not print.

SO WHAT IS HE, THEN?

Two hundred and sixty-six positions across three books. Three hundred and twenty verdicts, once you count the ones that split.

Thirteen have been overtaken.

That is the number this whole exercise produced, and it was not the number I expected. I went in expecting a list of things he got wrong. Thirteen out of three hundred and twenty is not a list. It is a rounding error attached to a career.

So the interesting question turned out not to be whether he is right. It is *how* he is right, because the pattern is unusually consistent and it is not the pattern you would guess.

The hedge is what saves him

This showed up in the first batch of research and then in every subsequent one, twelve times out of twelve, across fields with nothing in common.

Over and over, the claim in the body of a chapter is bounded exactly where a critic would want it bounded, and the bound is what makes it survive.

He said a gene contributes a little to sprinting and there may be hundreds more, giving the last word to the geneticist who discovered it. Thirteen years later that is the field's position, almost word for word. Without the hedge, the sentence he would be judged on is *a gene for speed*, and it would fail.

He said the muscle-fibre hypothesis about West African ancestry was unproven and untested in humans. It has still never been tested in a human being.

He said the slave-trade selection story was not supported by the DNA. A genome-wide scan of twenty-nine thousand people later found no signature of selection.

He said no sex test would ever classify cleanly, and three governing bodies have since spent a decade proving it, ending with policies that carry written exemptions for the exact condition his own opening story is about.

He said a self-experiment by one transgender runner showed a decline relative to her former self, and drew no conclusion about how she compared to other women.

That distinction is the entire argument the world has been having ever since, and he sidestepped it not by luck but because the evidence did not support crossing it.

The pattern is strong enough to state as a rule: the places where he declined to conclude are the places that have aged best. Not because refusing is safe, but because he refused where the evidence was absent, and absent evidence tends to stay absent.

The pull-quote is what sinks him

And in the same twelve batches of research, the overreaches turned up in the same structural position every time.

Not in the argument. At the *end of the section*, in the sentence built to be quoted, with the properly bounded version sitting one paragraph above it.

Total freedom is the enemy of creativity. The way to instantly make someone more creative is to force them to explore. Switchers are winners. The gap between men and women is widening.

Every one of those has a careful, defensible version a few lines earlier. Every one of them is stated in the slogan at a strength the paragraph does not support.

In one chapter of the newest book the same inference appears three times in three hundred words at three different strengths: the earlier results were *presumably* false positives, the rule change *caused* the shift, and *undoubtedly* there are ineffective products on the market as a result. Presumably, caused, undoubtedly, getting stronger as the section closes.

That is not carelessness. It is what happens to a hedge under narrative pressure, and it is remarkably consistent.

The practical instruction that falls out of it: quote the paragraph, never the line.

Three other habits, in increasing order of seriousness

He sometimes cites a paper for a detail and leaves out its headline.

The clearest case is the hinge of his newest book's closing argument. He argues that irreversible decisions leave people more satisfied, and cites a 2015 paper in the

footnote where he picks up a small closing detail about extreme maximisers. That paper's actual finding is that the effect reverses by personality type: satisficers prefer irreversibility, maximisers prefer the opposite. The reversal is by exactly the individual difference the chapter is about.

That happened four times in the two halves of that book, on claims that had already been corrected in print before he published.

He occasionally reports the wrong outcome variable. He describes a study as finding that narrow vocational education reverses into lower lifetime *earnings*. The study found a reversal in *employment*. Both are real findings; only one is that paper's.

And twice, a load-bearing claim has no citation at all.

The sentence that carries *Range* out of sport and into the world of work says he found a raft of studies showing inventors trade depth for breadth as their careers progress, and a nearly identical finding for artists. I checked the endnotes for that stretch of the introduction. Twelve other sentences nearby have notes. These two do not, and the literature that exists finds close to the reverse sequence.

The load-bearing claim of the music chapter, that breadth of training predicts breadth of transfer, is introduced as a classic research finding and is likewise unsourced.

There is a fourth category that looks like this and is not, and the distinction is worth keeping. The claim that specialists' contribution to invention peaked around 1985 rests on an unpublished algorithm run over ten million patents. But he attributes it in the text, by name, to the man who ran it. That is not concealment. It is honest attribution of something that cannot be checked. What he does not tell the reader is that nothing else supports it and the published literature points the other way.

The counter-examples, which are not optional

A portrait with only the faults in it would be a false portrait, and three things belong here.

He deleted his own best illustration. There is a famous study about children in a playground with and without a fence, which is quoted everywhere and is a perfect

demonstration of his newest book's thesis. It would have been the best story in it. He went looking for it, could not find it in any database, and printed the absence instead of the anecdote.

He publishes the caveat that weakens him. When two scientists told him the tendons could not be the whole story about the two high jumpers, he printed the refusal rather than burying it. When his subject denied that a genetic mutation had helped him win Olympic medals, he printed the denial and let the geneticist disagree, without adjudicating. When he reported an inventor's self-classification, he named the four roles that man played in the chapter without being asked to.

And one footnote in the newest book states his method outright, including an admission that he formed hypotheses after looking at data as a graduate student, that he used a tool to search deliberately for findings contradicting his own, and that no such effort is ever nearly perfect. That is more epistemic self-disclosure than most books of this kind contain anywhere.

The strangest thing in three books

In 2013 he diagnosed why single-gene studies of athletic ability were failing: the effects were too small to see in samples that size.

In 2026, writing about how science corrects itself, he reported that in genetics an entire field was basically wiped out, because researchers had been trawling databases for spurious correlations between traits and spots of DNA. The endnote adds, in a parenthesis: *this was also a topic in my first book.*

He described the destruction of the research programme a large part of his first book stood on, cited his own book, and did not say what the destruction did to it.

It is not dishonest. He was making a point about research methods, in a chapter about preregistration, and the fact was doing a different job there. Nothing obliged him to stop and audit himself.

But it is the thing I keep returning to, because it is the exact shape of what this book has been doing: two true statements, thirteen years apart, by the same person, that nobody had put next to each other.

What he actually is

Not a popular scientist who oversells, which is what I expected to find. Something rarer and more specific.

He is a reporter who is temperamentally uncomfortable concluding, working in a format that requires conclusions.

That tension explains almost everything in the pattern. The hedges are the reporter. The pull-quotes are the format. The refusals that aged best are the reporter winning; the slogans that failed are the format winning. The uncited hinges are what happens when a reporter needs a bridge between two subjects and does not have the study he needs, which is the one failure mode this book found that has no defence.

And it explains the thing that has surprised me most, which is how much of the work is *unchecked* rather than wrong. Eighty-six of the three hundred and twenty positions have never been tested by anybody. That is more than a quarter of everything a careful, well-sourced, best-selling writer has told several million people about how humans work, sitting in the literature with nobody having gone back to look.

He is not responsible for that. Nobody is, particularly. It is what a research system produces when the incentive is to publish something new rather than to check something old.

But it does mean that the most accurate one-line summary of an audit of David Epstein is not that he was right or wrong.

It is that he was careful, that carefulness is the whole reason he survived, and that a quarter of what he reported has been sitting there for over a decade with nobody bothering to find out.

ENDNOTES

Notes are grouped by chapter, and numbering restarts in each. They are attached only to claims that are surprising, contested, or worth checking for yourself. Everything else is in the working files.

Is There Such a Thing as a Sports Gene?

1. R. Border et al., "No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples," *American Journal of Psychiatry* 176, no. 5 (2019), doi:10.1176/appi.ajp.2018.18070881. Samples ranged from 62,000 to 443,000. The sport-specific version is T. Rankinen et al., "No Evidence of a Common DNA Variant Profile Specific to World Class Endurance Athletes," *PLOS ONE* 11, no. 1 (2016): e0147330, which examined 1,520 elite endurance athletes across eight cohorts and found a single surviving marker.
2. I. D. Papadimitriou et al., *BMC Genomics* 17 (2016): 285, doi:10.1186/s12864-016-2462-3. The figure is 0.92 percent, derived from 555 best times across 346 elite sprinters in ten countries. Kathryn North is a co-author.
3. M. Kamiya-Saito et al., "Association of the ACTN3 Genotype with Muscle Function: A Systematic Review and Meta-analysis," *International Journal of Sports Medicine* (2026), doi:10.1055/a-2799-1026, meta-analysing 53 articles.
4. *Inside the Box*, chapter 3, endnote 16, citing C. Wilson, "Failure to Replicate," *New Scientist*, April 2022.

Does the Same Training Work on Everyone?

1. P. J. Williamson, G. Atkinson, A. M. Batterham, "Inter-Individual Responses of Maximal Oxygen Uptake to Exercise Training: A Critical Review," *Sports Medicine* 47, no. 8 (2017): 1501-1513, doi:10.1007/s40279-017-0680-8, reanalysing D. Prud'homme et al., *Medicine & Science in Sports & Exercise* 16, no. 5 (1984): 489-493.
2. M. Del Giudice et al., *Journal of Science and Medicine in Sport* 23, no. 1 (2020): 94-99, doi:10.1016/j.jsams.2019.09.007.
3. R. J. Shephard, T. Rankinen, C. Bouchard, "Test-retest errors and the apparent heterogeneity of training response," *European Journal of Applied Physiology* 91, no. 2-3 (2004): 199-203, doi:10.1007/s00421-003-0990-4. The observed spread in training response was 202 millilitres per minute; after subtracting test-retest error, 123 to 138.
4. S. J. R. Meyler et al., "Changes in Cardiorespiratory Fitness Following Exercise Training Prescribed Relative to Traditional Intensity Anchors and Physiological Thresholds," *Sports*

Medicine (2024), doi:10.1007/s40279-024-02125-x, systematic review and meta-analysis of 42 studies and 1,544 participants.

Can You See It Coming?

1. D. T. Y. Mann, A. M. Williams, P. Ward, C. M. Janelle, "Perceptual-Cognitive Expertise in Sport: A Meta-Analysis," *Journal of Sport and Exercise Psychology* 29, no. 4 (2007), doi:10.1123/jsep.29.4.457. Forty-two studies, 388 effect sizes.
2. M. W. Voss, A. F. Kramer, C. Basak, R. S. Prakash, B. Roberts, "Are expert athletes 'expert' in the cognitive laboratory?", *Applied Cognitive Psychology* (2010), doi:10.1002/acp.1588; and H.-E. Scharfen, D. Memmert, "Measurement of cognitive functions in experts and elite athletes: A meta-analytic review," *Applied Cognitive Psychology* (2019), doi:10.1002/acp.3526.
3. T. Filippini et al., "Professional baseball players demonstrate superior dynamic visual acuity and differences in vestibulo-ocular reflex performance, compared to similarly aged healthy controls," *BMC Sports Science, Medicine and Rehabilitation* (2025), doi:10.1186/s13102-025-01405-x.
4. J. G. Tiernan et al., "Visual Perception in Elite, Fast-Paced Ball-Sport Athletes: A Systematic Review and Meta-Analysis," preprint, doi:10.31234/osf.io/b5jzw_v1. Twenty-six studies. Elite versus non-athlete, Hedges' $g = 0.84$; elite versus sub-elite, $g = 0.12$, not significant.

Why Do the Bodies Keep Getting Stranger?

1. K. Norton, T. Olds, "Morphological evolution of athletes over the 20th century: causes and consequences," *Sports Medicine* 31, no. 11 (2001), doi:10.2165/00007256-200131110-00001. The finding that athlete morphology outpaced the general population's secular trend is theirs. The "Big Bang" metaphor is David's.
2. K. Emig, J. Adam, "Evolution of world running record performances for men and women: physiological characteristics," *Frontiers in Physiology* (2024), doi:10.3389/fphys.2024.1372092.
3. The five-to-one muscle-to-bone ratio is attributed in *The Sports Gene* to Francis Holway's measurements and I could not locate it in any published source. Reported here as an unpublished figure attributed to a named researcher.
4. The seventeen-percent seven-footer estimate and the arm-span ratio are both David's own analyses, described as such in the book. Neither has been independently published.

Does Ancestry Explain Who Wins?

1. V. O. Onywera et al. (2006), questionnaires administered to 404 elite Kenyan runners and 87 controls. The same gradient appears for region, ethnicity and language group.

2. H. B. Larsen, A. W. Sheel, "The Kenyan runners," *Scandinavian Journal of Medicine & Science in Sports* (2015).
3. H. B. Larsen et al., *Experimental Physiology* (2026). Twelve Kenyan elite, 29 Kenyan sub-elite, 20 Danish elite, 37 Danish sub-elite adult runners, plus 46 untrained Kenyan and 30 untrained Danish adolescents.
4. M. Deason et al., "Importance of mitochondrial haplotypes and maternal lineage in sprint performance among individuals of West African ancestry," *Scandinavian Journal of Medicine & Science in Sports* 22, no. 2 (2012): 217-223, doi:10.1111/j.1600-0838.2010.01289.x. The Jamaican comparison was null at $P = 0.551$; the African-American maternal-lineage comparison was significant at $P < 0.001$.
5. G. Bhatia et al., "Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture," *American Journal of Human Genetics* (2014), doi:10.1016/j.ajhg.2014.08.011.
6. L. D. Weeks et al., "Sickle cell trait does not cause 'sickle cell crisis' leading to exertion-related death: a systematic review," *Blood* 145, no. 13 (2025): 1345-1352, doi:10.1182/blood.2024026899.

Who Counts as a Woman?

1. V. Thibault et al. (2010), analysing 82 events, found stabilisation after 1983 at a mean gap of 10.0 percent, with running at 10.7, jumps at 17.5 and swimming at 8.9.
2. S. Bermon et al., "Serum Androgen Levels in Elite Female Athletes," *Journal of Clinical Endocrinology & Metabolism* (2014), doi:10.1210/jc.2014-1391. Screening of 849 elite female athletes.
3. D. J. Handelsman, S. Bermon, "Serum androgen profile of elite athletes," *Journal of Clinical Endocrinology & Metabolism* (2026), doi:10.1210/clinem/dgag196. N = 1,689.
4. J. Harper et al. (2025), ten transgender women runners, mean decline 14.6 percent over approximately 31 months.
5. The three-standard-deviation throwing figure is attributed in *The Sports Gene* to Thomas and French (1985). Full text was not obtainable and the figure is reported here as unverified.

What Is Genetic Testing Actually For?

1. S. R. Ommen et al., 2024 AHA/ACC/AMSSM/HRS/PACES/SCMR guideline for the management of hypertrophic cardiomyopathy, doi:10.1161/CIR.0000000000001250.
2. R. Hespe et al., *Journal of the American College of Cardiology* (2025), doi:10.1016/j.jacc.2024.12.010. Reappraisal of gene-disease validity by the ClinGen expert panel.

3. From the SHaRe registry: 12,187 patients, 8,054 genotyped, 1,606 unique variants, 17 percent reclassified on systematic reuration.
4. J. H. Kim et al., "Clinical Considerations for Competitive Sports Participation for Athletes With Cardiovascular Abnormalities," *Journal of the American College of Cardiology* (2025), doi:10.1016/j.jacc.2024.12.025.
5. R. Lampert, K. G. Harmon, "Sudden Cardiac Arrest in Athletes," *New England Journal of Medicine* (2026), doi:10.1056/NEJMra2312555.
6. Wilson et al. (2026), systematic review across 33 studies, doi:10.1017/cjn.2026.10581.

When Does Experience Actually Make You Better?

1. R. M. Hogarth, T. Lejarraga, E. Soyer (2015) give the cleanest statement of the distinction: kind environments involve close matches between the information available while learning and the information available while predicting; wicked environments involve mismatches. The originating paper is D. Kahneman and G. Klein, "Conditions for intuitive expertise: A failure to disagree," *American Psychologist* (2009).
2. F. Gobet, H. A. Simon (1996, 1998) re-ran the original chess recall experiments, confirmed chunking, found masters using larger chunks than the 1973 study reported, and extended the theory to templates in order to account for skilled recall of random and distorted positions.

What Makes Learning Stick?

1. C. Brunmair, T. Richter, meta-analysis of interleaved practice: $g = 0.67$ for paintings, $g = 0.34$ for mathematics, $g = -0.39$ for word learning.
2. S. E. Carrell, J. E. West (2010). The study's own abstract states that professors who excel at promoting contemporaneous achievement teach in ways that improve their evaluations but harm the follow-on achievement of their students in more advanced classes.
3. D. H. Bailey, G. J. Duncan, F. Cunha, B. R. Foorman, D. S. Yeager (2020), concluding that fadeout is a substantive phenomenon rather than a measurement artefact.
4. E. R. Hart, D. H. Bailey, S. Luo, P. Sengupta et al., "Fadeout and persistence of intervention impacts on social-emotional and cognitive skills in children and adolescents," *Psychological Bulletin* (2024), doi:10.1037/bul0000450. Eighty-six randomised trials, 56,662 participants.

Is Grit a Thing?

1. M. Credé, M. C. Tynan, P. D. Harms, "Much ado about grit: A meta-analytic synthesis of the grit literature," *Journal of Personality and Social Psychology* (2017), doi:10.1037/pspp0000102.

Perseverance of effort with academic performance at $p = .26$, consistency of interest at $p = .10$.

2. Lam and Zhou (2022), overall $r = .19$, perseverance $r = .21$, consistency $r = .08$, across more than 285,000 participants, finding no significant moderators of the overall grit-achievement link.
3. A. L. Duckworth, P. D. Quinn, M. E. P. Seligman et al., "Cognitive and noncognitive predictors of success," *PNAS* 116, no. 47 (2019), doi:10.1073/pnas.1910510116, N = 11,258 West Point cadets, is the source of the negative correlations described above.

Can Anyone Forecast Anything?

1. W. Chang, E. Chen, B. Mellers, P. Tetlock (2016), reporting that a cognitive-debiasing module lasting less than one hour consistently improved Brier scores by 6 to 11 percent over a randomised control condition, within a four-year government-sponsored tournament.
2. M. N. Stagnaro, B. M. Tappin, D. G. Rand, *PNAS* (2023), failing to replicate the motivated numeracy effect in a nationally representative US probability sample.
3. S. R. Dalal, E. B. Fowlkes, B. Hoadley, *Journal of the American Statistical Association* (1989), using all 23 pre-accident launches and finding strong statistical evidence of a temperature effect on O-ring thermal distress, with a probabilistic risk assessment at the launch temperature.
4. E. M. Anicich, R. I. Swaab, A. D. Galinsky, *PNAS* (2015), analysing 30,625 Himalayan climbers from 56 countries across 5,104 expeditions.
5. The organisational culture study is J. C. Smart, E. P. St. John, "Organizational Culture and Effectiveness in Higher Education: A Test of the 'Culture Type' and 'Strong Culture' Hypotheses," *Educational Evaluation and Policy Analysis* (1996), doi:10.3102/01623737018003219. Full text not obtained; the characterisation above is from the published abstract.

Should You Keep Your Options Open?

1. O. Malamud, "Discovering One's Talent: Learning from Academic Specialization," *ILR Review* 64, no. 2 (2011), doi:10.1177/001979391106400209, and "Breadth versus Depth: The Timing of Specialization in Higher Education," *LABOUR* 24, no. 4 (2010), doi:10.1111/j.1467-9914.2010.00489.x.
2. E. A. Hanushek, G. Schwerdt, L. Woessmann, L. Zhang, "General Education, Vocational Education, and Labor-Market Outcomes over the Lifecycle," *Journal of Human Resources* 52, no. 1 (2017), doi:10.3368/jhr.52.1.0415-7074r. The outcome is employment across the lifecycle.

3. W. Bleidorn, T. Schwaba, A. Zheng, C. J. Hopwood, *Psychological Bulletin* 148, nos. 7-8 (2022), doi:10.1037/bul0000365. Rank-order stability across 189 samples, $N = 178,503$; mean-level change across 276 samples, $N = 242,542$.
4. T. W. Watts, G. J. Duncan, H. Quan, *Psychological Science* 29, no. 7 (2018), doi:10.1177/0956797618761661.
5. S. D. Levitt, "Heads or Tails: The Impact of a Coin Toss on Major Life Decisions and Subsequent Happiness," *Review of Economic Studies* 88, no. 1 (2021), doi:10.1093/restud/rdaa016. Approximately 20,000 recruited, 2,186 with qualifying decisions.

Does Breadth Actually Pay?

1. L. B. Jeppesen, K. R. Lakhani (2010), studying 166 problems and more than 12,000 solvers on a broadcast-search platform. Note also that distant solvers produce ideas rated *less feasible*, a documented cost the book does not report.
2. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, "Atypical Combinations and Scientific Impact," *Science* (2013), doi:10.1126/science.1240474, covering 17.9 million papers; and J. Wang, R. Veugelers, P. Stephan, *Research Policy* (2017), doi:10.1016/j.respol.2017.06.006, covering approximately one million papers. Extended by Hill et al., *Nature* (2025), doi:10.1038/s41586-025-09048-1.
3. L. Liu et al., *Nature Communications* (2021), doi:10.1038/s41467-021-25477-8, finding an explore-then-exploit sequence with the narrowing coinciding with peak impact.
4. F. Teodoridis, M. Bikard, K. Vakili, *Administrative Science Quarterly* (2019), doi:10.1177/0001839218793384.
5. The polymath study is E. Melero, N. Palomeras, "The Renaissance Man is not dead! The role of generalists in teams of inventors," *Research Policy* (2015), doi:10.1016/j.respol.2014.07.005. I was unable to obtain its full text, so its sample figures are not verified here. The three-way split described above is from the abstract of the separate corporate study the chapter's protagonist co-authored.

Do Constraints Make You Better?

1. A. Camuffo, A. Cordova, A. Gambardella, C. Spina, "A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial," *Management Science* 66, no. 2 (2020): 564-586, doi:10.1287/mnsc.2018.3249. The replication is A. Camuffo, A. Gambardella, D. Messinese, E. Novelli, E. Paolucci et al., "A scientific approach to entrepreneurial decision-making: Large-scale replication and extension," *Strategic Management Journal* (2024), doi:10.1002/smj.3580.
2. The uncited meta-analysis of the book's own thesis is Damadzic et al. (2022), *Journal of Organizational Behavior*, covering 111 studies. The discovery-learning meta-analysis is L.

Alfieri et al. (2011), 164 studies, finding unassisted discovery worse than explicit instruction and enhanced discovery better than both.

3. Several figures in this book could not be independently sourced and are reported here as unverified: the megaproject base rates, which come from a trade book rather than a paper; a hundred-million-fold figure attributed to a business writer; and the sample size of the 2010 creativity study underpinning one of its stronger claims.

What Should You Settle For?

1. J. Holt-Lunstad et al., meta-analysis of 148 studies and 308,849 participants, reporting an odds ratio of 1.50 for survival with stronger social relationships, stable across age, sex, baseline health status, cause of death and follow-up period.
2. B. Scheibehenne, R. Greifeneder, P. M. Todd (2010), meta-analysis of the choice-overload literature finding a mean effect near zero; and A. Chernev, U. Böckenholt, J. Goodman, restoring an effect under four moderators while conceding the open question.
3. N. N. Cheek, B. Schwartz (2016), counting eleven competing measures and describing the literature as befuddling and contradictory. Schwartz is a co-author of the original Maximization Scale.
4. R. Shiner, "Maximizers, Satisficers, and Their Satisfaction With and Preferences for Reversible Versus Irreversible Decisions," *Social Psychological and Personality Science* (2015), doi:10.1177/1948550615595271.
5. Several figures in the closing section could not be independently verified and are not reproduced here, including a widely quoted arc of declining attention spans whose later values have no peer-reviewed source.

So What Is He, Then?

1. The counts throughout: 266 positions extracted from *The Sports Gene* (2013), *Range* (2019) and *Inside the Box* (2026); 320 verdicts after splitting positions that resolved differently across their clauses; 13 SUPERSEDED, 86 UNTESTED, 85 MAINSTREAM, 73 CONTESTED, 47 VINDICATED, 16 with no testable empirical content.
2. Positions his own later books withdrew or narrowed were excluded from the audit before it began, and do not appear in these counts. The largest such exclusion is the afterword to the paperback edition of *Range*, in which he reclassifies several of his own set-pieces as counterweights to an availability bias rather than as evidence, concedes that early specialisation is legitimate as one path among several, and withdraws the quarrel with the grit literature.