



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/246306/>

Version: Accepted Version

Article:

Bilalic, M., Vaci, N., Hönekopp, J. et al. (Accepted: 2026) Studying the best can mislead: selection bias in expertise research. *Perspectives on Psychological Science*. ISSN: 1745-6916 (In Press)

© 2026 The Author(s). This is an author-produced version of a paper accepted for publication in *Perspectives on Psychological Science*. Uploaded in accordance with the publisher's self-archiving policy. Article available under the terms of the CC-BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Studying the Best Can Mislead: Selection Bias in Expertise Research

Merim Bilalić ^{1*}

Nemanja Vaci ²

Johannes Hönekopp ¹

Thomas V. Pollet ¹

Affiliations:

¹ Northumbria University, UK

² University of Sheffield, UK

Author Note

Merim Bilalić <https://orcid.org/0000-0001-9523-6734>

Nemanja Vaci <https://orcid.org/0000-0002-8094-0902>

Johannes Hönekopp <https://orcid.org/0000-0002-0613-6801>

Thomas Pollet <https://orcid.org/0000-0002-4210-935X>

We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Merim Bilalić, School of Psychology, Northumbria University, Newcastle upon Tyne, UK. Email: merim.bilalic@northumbria.ac.uk

Data and Code for Analyses:

https://osf.io/9fas2/overview?view_only=f38e297d70af470c8a65cedc20db1567

Abstract

What we think we know about how people reach exceptional performance could be distorted by who gets studied. Studying elite performers has been highly informative for understanding the cognitive mechanisms of expertise but might be less suitable for explaining how expertise develops. Elite samples are filtered on the very success researchers want to explain. Predictors that appear beneficial early in development often weaken or even reverse among world-class adults. These findings are commonly interpreted as evidence that the route to excellence changes at higher performance levels. Here we demonstrate that they can also arise from selection alone. Outstanding adult performance is rare, depends on many factors, and can be reached in more than one way. Focusing exclusively on those who succeed can create misleading trade-offs, make statistical relationships look smaller, and make weak positive and weak negative effects look the same. Simulations and longitudinal data from chess and athletics illustrate these patterns and show that they depend on who remains in the sample and how narrowly elite performance is defined. We offer practical principles for studying elite performers without mistaking associations within selected samples for evidence about how excellence develops.

Keywords: expertise, outstanding performance, selection bias, collider bias, talent identification, deliberate practice, causal inference, directed acyclic graphs (DAGs)

Introduction

Watching outstanding performers invites two questions: how do they achieve such extraordinary performances, and how do people get there in the first place? Researchers have often tried to answer both by studying those who have already succeeded: chess grandmasters, world-class athletes, top musicians, and eminent scientists. This strategy has been highly productive for understanding the structure of expert cognition and performance. However, it may be less reliable for explaining how expertise develops, because these studies only include people who have already reached the outcome it is trying to explain. Once researchers restrict attention to those who have “made it,” they are no longer observing a neutral sample of developing performers. Instead, they are studying a group that has passed through one or more selection processes and was filtered on outcomes shaped by many factors acting together. As a result, the associations seen in elite samples can differ sharply from those in the broader population of aspirants. They can weaken, disappear, or even reverse. Here, we argue that this is not a peripheral methodological nuisance but rather a central inferential problem in research on outstanding performance.

The attraction to elite samples is understandable. From the earliest days of scientific psychology, outstanding performers have served as a window into the mind (Binet, 1894). Classic work on chess, for example, showed that experts do not necessarily think faster or possess greater general ability (Chase & Simon, 1973; de Groot, 1946/1978). Instead, they rely on domain-specific knowledge structures that shape perception, guide attention, support anticipation, and allow rapid recognition of what matters (Bilalić, 2017, 2018; Costantino et al., in press; Gobet & Simon, 1996; Gobet et al., 2001). For questions about the cognitive structure of expertise, elite performers were therefore an especially revealing

population (McCarthy, 1990; Simon & Chase, 1973). The difficulty arose when the same strategy was extended from describing *what* expertise looks like to explaining *how* it develops (Ericsson et al., 1993; Hambrick et al., 2014; Ullén et al., 2016).

Across several major debates on development of expertise, the same pattern occurs: what helps early is thought to stop helping, or become harmful, at the highest levels. This pattern, which is often interpreted in developmental terms, is the target of the present paper.

Two influential controversies illustrate the pattern. The first concerns deliberate practice. The deliberate-practice framework proposes that highly structured, effortful training is the main factor behind exceptional achievement (Ericsson et al., 1993). That claim has been debated extensively (Campitelli & Gobet, 2011; Ericsson & Harwell, 2019; Hambrick et al., 2020; Southwick et al., in press), but one empirical result recurs across the debate: practice predicts performance in broad samples, then appears to matter less as samples become more selective. Meta-analyses show substantial associations in domains such as chess, music, and sport, but much smaller ones in more filtered domains such as education and professions (Macnamara et al., 2014; Macnamara et al., 2016). Within sport, the association is markedly weaker among elite athletes than in mixed-performance samples (Macnamara et al., 2016), and in some elite samples it even appears to reverse (Güllich et al., 2022, 2025).

The second controversy concerns specialization versus diversification debate. Recent meta-analyses report a striking contrast between youth and adult elite performers. Among juniors, earlier specialization, more main-sport practice, and faster early progress are associated with higher achievement. Among world-class adults, the profile appears reversed: later specialization, more multisport experience, less early main-sport practice,

and slower early progress are more common than among national-class athletes (Barth et al., 2022; Güllich & Barth, 2024; Güllich et al., 2022, 2025). These results have been interpreted not only as evidence that early and late excellence follow qualitatively different developmental pathways, but also as evidence that early specialization may be counterproductive for later world-class attainment (Güllich et al., 2022, 2025).

These debates matter well beyond theory. Claims about what predicts outstanding performance are routinely translated into decisions about talent identification, academy entry, youth training design, educational streaming, and institutional investment (Ford et al., 2011; Rees et al., 2016; Till & Baker, 2020). If selection can produce these patterns, then the developmental conclusions drawn from them may not hold. The practical consequences could be substantial: one may select the wrong individuals, invest in the wrong developmental practices, and treat preventable dropout or delayed development as evidence of lower potential.

The selection problem itself is not new. Related forms of collider bias, survivorship bias, attrition, and sample selection have long been studied in epidemiology, sociology, economics, and, more recently, psychology (e.g., Berkson, 1946; Elwert & Winship, 2014; Heckman, 1979; Hernán et al., 2004; Munafò et al., 2018; Rohrer, 2018; Wysocki et al., 2022). Here we demonstrate why this logic is especially consequential for research on expertise. Elite samples are not only small or unrepresentative – they are samples of people who have already survived participation, opportunity, retention, and performance thresholds. Within such samples, patterns that look like developmental reversals could arise even when no reversal in the underlying developmental process has occurred. The central question is therefore not whether early predictors still forecast success among elites, but what those within-elite patterns can and cannot tell us about how excellence develops.

We develop our argument in four steps. First, we illustrate the selection mechanism with Directed Acyclic Graphs (DAGs) and a simple chess example. Second, we use simulations to ask whether selection alone can reproduce the key pattern in the literature: a positive early predictor becoming weak or negative among elites. Third, we examine real-world longitudinal data from chess and athletics to test whether the same threshold-dependent attenuation appears outside the simulation. Finally, we outline practical diagnostics and reporting defaults for studying elite performers without treating selected-sample patterns as evidence about development.

Selection Bias in Elite Samples

Elite samples¹ are not merely unrepresentative. Their deeper inferential problem is that membership in such samples is itself a consequence of success, and success is usually produced by multiple contributing factors. This is a form of selection bias. One important mechanism is collider bias. When researchers analyze only the most successful practitioners, they are effectively conditioning on success, that is, restricting attention to people who reached an outcome produced by several causes at once. This can make those causes appear negatively related within the selected sample even if they were originally unrelated, or even positively related, in the broader population (Berkson, 1946; Elwert & Winship, 2014; Munafò et al., 2018; Pearl, 2009; Ren & Loh, 2024; Richardson et al., 2019; Wysocki et al., 2022).

A familiar and intuitive example comes from professional basketball: height and shooting skill both help players reach the NBA, yet within the NBA height and free-throw

¹ We use "elite" operationally rather than as a fixed category. In the conceptual sections it refers broadly to performers selected for high achievement. In the simulations and analyses it refers to an upper tail of adult performance defined by a cutoff. As we will demonstrate, the size and even the sign of within-sample associations depend on where that cutoff is set.

accuracy may be negatively associated. This does not mean that being taller makes players worse shooters. It means that shorter players must compensate with exceptional shooting to reach the same selected group. Taller players, on the other hand, can qualify despite weaker shooting. The resulting trade-off is descriptively real within the elite sample but causally misleading (Cole et al., 2010; Elwert & Winship, 2014).

The same logic can be made explicit with DAGs (Grosz et al., 2020; Kunicki et al., 2023; Pearl, 2009; Rohrer, 2018) and shown directly in developmental data. Figure 1 illustrates the structure using data from 57 eleven-year-old chess players tested on intelligence, accumulated practice, and chess skill (Bilalić, McLeod, & Gobet, 2007; see also Bilalić, 2017). In the full sample, both intelligence and practice predict skill ($r = .55$ and $r = .90$, respectively). They are also positively correlated with each other ($r = .45$).

The DAG captures this structure (Figure 1A): two contributing causes (intelligence and practice) jointly produce the outcome (skill). The picture changes once we select (condition) on the outcome. Focusing on the top 20 children by skill produces three distortions.

First, intelligence and practice become negatively associated within the selected group (Figure 1, Panel B). Among the best young players, those who got there through high intelligence tended to have less practice, and vice versa. This apparent trade-off was absent in the full sample and emerged only after the top 20 players were selected.

Second, the strong positive link between intelligence and skill shrinks and can even flip sign as the selection tightens (Figure 1, Panel C).

Third, even the near-perfect practice–skill correlation is attenuated toward zero among the very best (Figure 1, Panel D). The same predictor that explains almost everything in the full sample explains approximately a third less in the top-tail.

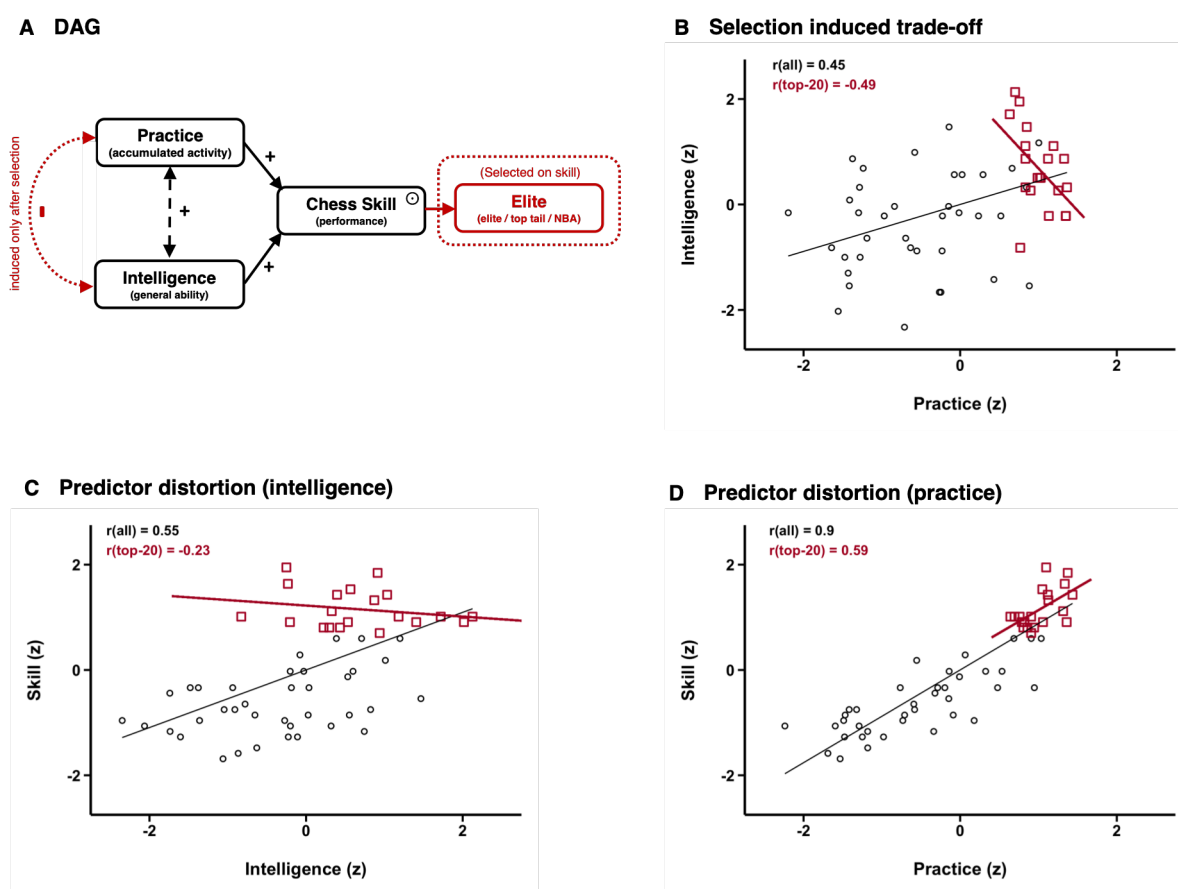


Figure 1. Selection on chess skill changes associations among predictors and outcomes.

Data are from 57 eleven-year-old chess players tested on intelligence, accumulated practice, and chess skill (Bilalić, McLeod, & Gobet, 2007). **(A)** Directed acyclic graph: intelligence and practice both contribute to chess skill, and the elite subgroup is defined by selecting the top tail of skill. **(B)** Association between practice and intelligence in the full sample and among the top 20 players. **(C)** Association between intelligence and chess skill in the full sample and among the top 20 players. **(D)** Association between practice and chess skill in the full sample and among the top 20 players. Gray lines show linear fits in the full sample; red lines show linear fits within the selected subgroup.

Three further sources of distortion compound the picture. The first is range restriction (Sackett & Yang, 2000; Thorndike, 1949; Vaci et al., 2014). When we pick the top tail, the remaining spread in the outcome, and in any predictor correlated with it, is much smaller than in the full population. Correlations computed in this compressed range are attenuated, sometimes drastically. Once the correlation is weakened, it lies closer to zero,

which makes random sign changes more likely from sample to sample. Range restriction does not itself create a compensatory mechanism – it only weakens the association.

Once the association has been weakened, sampling variability becomes more consequential. Elite samples are often small, and estimates based on compressed variation can move substantially from sample to sample. In such settings, the observed association could approach zero or even reverse sign by chance.

Finally, there is also sequential selection (Elwert & Winship, 2014; Hernán et al., 2004; Howe et al., 2016). Before anyone defines "elite," the adult sample has already been filtered by those who continued to participate – who kept playing, who stayed in the database, who did not drop out (Rohrer & Murayama, 2023). Before an elite cutoff is applied, those who remain into adulthood may already differ systematically from the original cohort. Elite samples therefore reflect two stages of selection: who remained and who reached the top.

Table 1 summarizes these four sources of distortion. The first three can change the association expected even in a very large sample. Sampling variability then determines how far a particular estimate moves around that expectation. A within-elite reversal can therefore arise from the selection process itself and should not, on its own, be treated as evidence for a developmental process.

Table 1
Four Sources of Distortion in Within-Elite Associations

Source of distortion	What it does	Consequence within elite samples
Collider bias	Conditioning on success, which has several causes, induces dependence among those causes	The causes appear negatively related within elites – a trade-off absent in the population

SELECTION BIAS IN EXPERTISE RESEARCH

Source of distortion	What it does	Consequence within elite samples
Range restriction	Selecting the top tail compresses variance in the outcome and its correlates	Predictor–outcome associations shrink toward zero; no compensatory trade-off is implied
Sequential selection / survivorship	Dropout filters the cohort before “elite” is defined	The adult sample no longer represents the original cohort; distortion accrues before any elite cutoff
Sampling variability	The compressed relationship is estimated from a small elite sample	At realistic sample sizes the observed sign is unstable and can reverse by chance

These distortions are easiest to diagnose when the relevant contributors are measured, as in the chess illustration. In most developmental settings, however, important contributors to adult performance, such as maturation, coaching quality, opportunity, family resources, and psychosocial support, are partly or wholly unmeasured. The same selection logic may therefore operate even when the compensatory routes into elite status cannot be observed directly.

Computational Simulation

The simulation follows a simplified path from the initial population, through continued participation, to the elite group. It asks two questions: 1) can selection make an early predictor that genuinely helps later performance appear weak or harmful among elites, and if so, 2) can data from elites alone distinguish that pattern from one in which the predictor is genuinely harmful? The aim is not to reproduce a particular developmental theory. If positive and negative developmental relationships produce the same pattern among elites, evidence from outside the elite sample is needed to tell them apart.

Our model has two gates. First, some young performers remain visible into adulthood and others do not. Second, among those who remain, only a tiny top fraction count as elite, with cutoffs such as the top 50 or top 100 defined within simulated cohorts of $N = 100,000$ (see, Mitchell et al., 2024). Early specialization helps youth performance in the model. Adult performance, however, depends not only on youth performance but also on later-emerging influences that youth results do not fully capture. This two-stage process, survival followed by elite selection, provides the setting for the test.

Figure 2A shows how the variables are related in the simulation. Two independent factors produce youth performance. One is specialization (S): early start, volume of main-domain practice, and sport focus. The other is talent (T): predispositions that support learning and performance. Specialization contributes more strongly to youth performance than talent does. This reflects both prior evidence on the importance of early accumulated domain-specific exposure and a deliberately conservative modeling choice². Youth performance then determines passage through the first gate: weaker youth performers are more likely to disappear from the pipeline. As a result, those who remain into adulthood tend to have performed better in youth and to have specialized more (see Figure 2B).

Adult performance (A) depends on youth performance (Y) and on additional influences (U) that emerge later, such as maturation, coaching, opportunity, or psychosocial support³. We vary the strength of the youth-to-adult carry-over, that is, the extent to which doing better in youth tends to translate into doing better in adulthood before later-

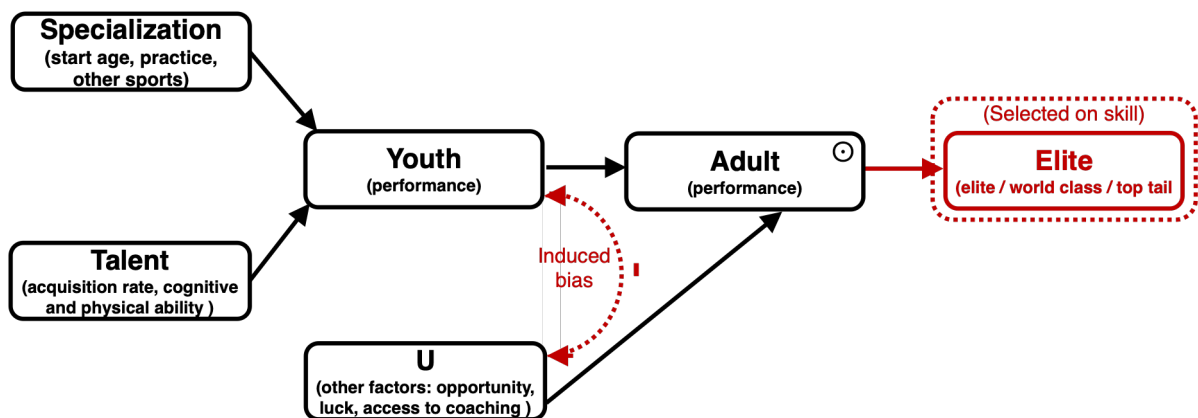
² We deliberately give specialization a strong role in youth performance. This is consistent with the chess example, where accumulated practice is more strongly related to youth skill than intelligence is. If selection can nevertheless remove the specialization–performance association among elites, the result cannot be attributed to giving specialization too little influence in the simulation.

³ The model does not attempt to represent the dynamic interactions among these later-emerging U factors over time. Instead, we use them here to test whether selection on adult performance can distort the observed relation between early predictors and adult outcomes.

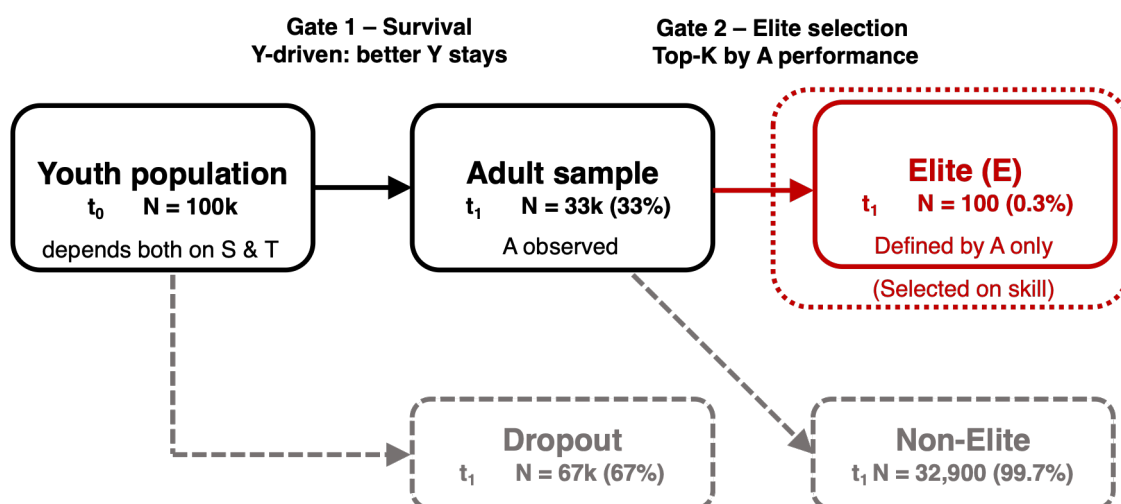
SELECTION BIAS IN EXPERTISE RESEARCH

emerging influences are taken into account. In the main simulation, neither specialization nor talent has a direct path to adult performance – both influence adulthood only through their contribution to youth performance (we relax this assumption in the Supplemental Materials, SM, Section 3.4).

A Causal structure



B Selection gates (youth → adult → elite)



C Why selection can make the specialization–adult link look negative

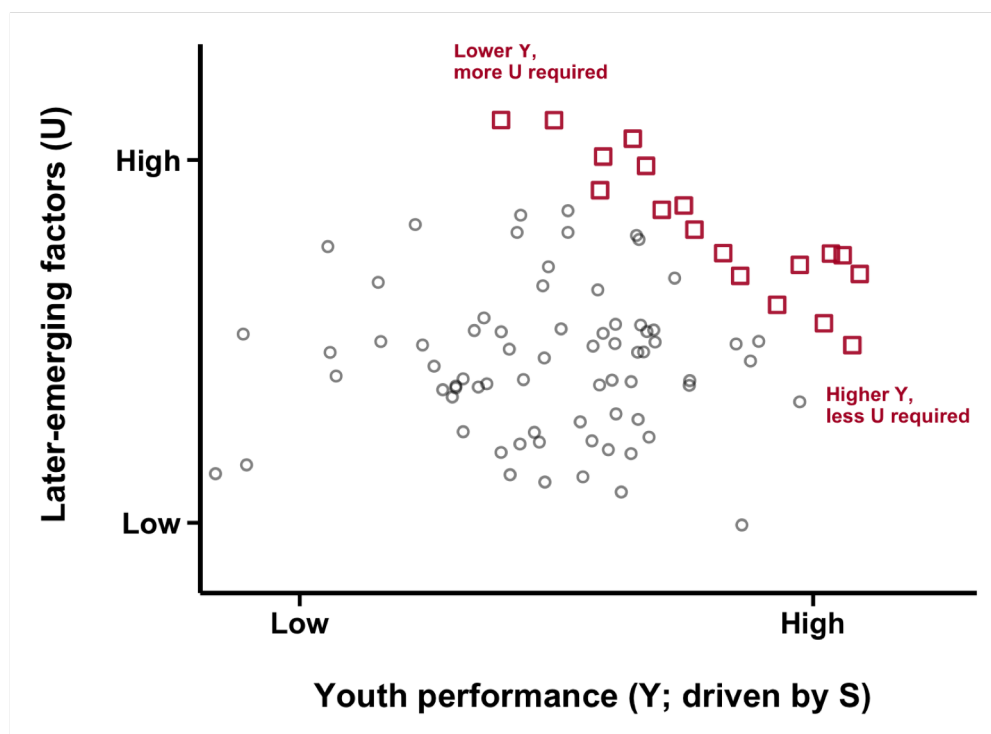


Figure 2. Causal structure and selection mechanism. (A) **Causal structure.** Specialization (S) and talent (T) produce youth performance (Y). Adult performance (A) depends on Y and unmeasured factors (U). There is no direct $S \rightarrow A$ path in this model: specialization reaches A only through Y. Selecting on elite status (top tail of A) induces a negative dependency between Y and U (dashed arrow), which propagates to an apparent negative S–A association. (B) **Selection gates (youth $\rightarrow$ adult $\rightarrow$ elite).** A youth performance-driven survival gate produces the adult sample, after which elite status is defined by adult performance. Gray boxes indicate those lost at each gate. (C) **Why selection can make the specialization–adult link look negative.** Black circles represent the broader adult sample, and red squares the elite group. Within elites, lower youth performance (Y) requires higher later-emerging influences (U), whereas higher Y requires less U. Because specialization (S) contributes to Y, this trade-off can make the within-elite S–A association appear weak or negative.

The elite group is then formed by selecting the highest-performing adults among those who survive the first gate. The number selected determines how exclusive the elite group is. Because adult performance depends on both youth performance and later-emerging influences, people can reach the same adult level through different combinations of the two. This creates a compensatory trade-off within the elite group: those who arrive

through stronger youth performance need less later advantage, and those who arrive through stronger later advantage need less youth performance (Figure 2C). Because specialization is a major component of youth performance, that trade-off can pull its within-elite association with adult performance toward zero, or below.

We examine the specialization–adult-performance association at three levels: in the full population, among adult survivors after Gate 1, and within the elite subset after Gate 2. The simulations vary only the strength of the carry-over from youth to adult performance:

Scenario A: strong positive youth-to-adult carry-over (+0.50).

Scenario B: weak positive youth-to-adult carry-over (+0.20).

Scenario C: weak negative youth-to-adult carry-over (−0.20).

The critical comparison is between Scenarios B and C. They have the same absolute size but opposite signs. This is the key contrast for judging whether elite-only data can discriminate between the two accounts (full parameter values are reported in the SM, Section 3).

In Scenario A, the population association is strong and remains visible even after selection (Figure 3). Figure 3A plots the within-elite correlation as progressively more adults are included, moving from the most selective group toward all adult survivors. The specialization–adult-performance correlation is clearly positive in the population and, although attenuated within elites, stays positive at moderate cutoffs. Only at the most extreme cutoffs does it become unstable enough to reverse to a negative sign.

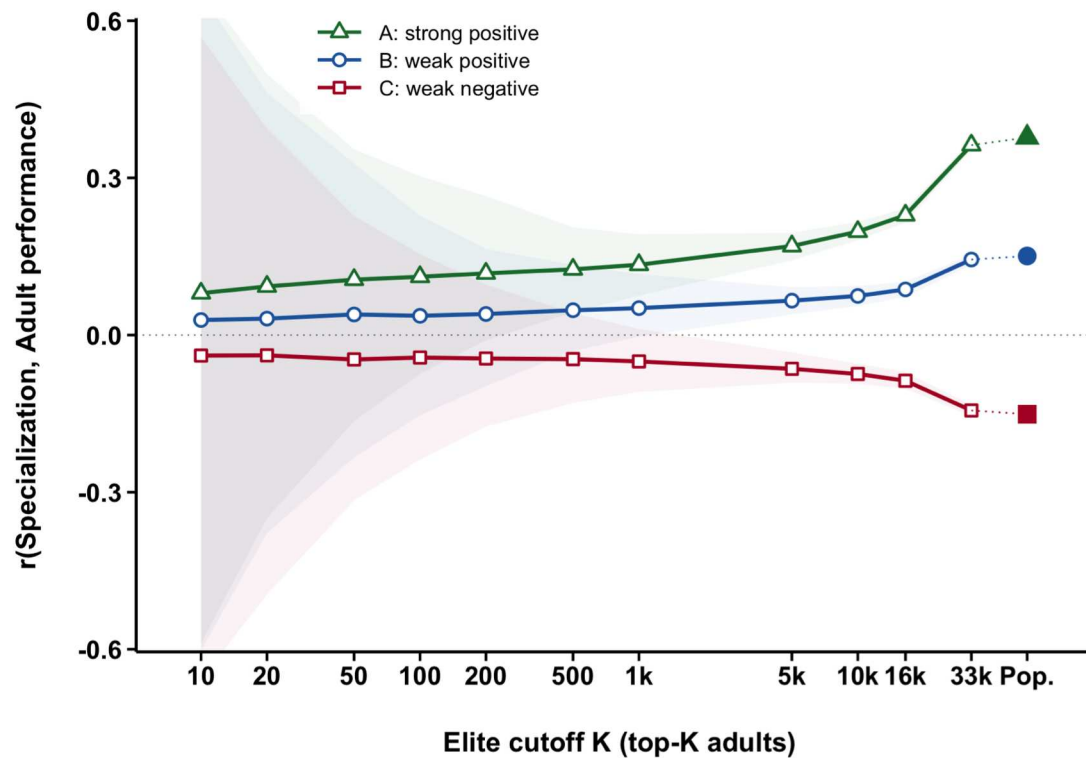
Scenarios B and C are unmistakably different in the population: specialization is positively related to adult performance in Scenario B and negatively related in Scenario C.

But as the elite cutoff tightens, both estimates are squeezed toward zero. Among the top 100, the two scenarios are almost indistinguishable – the 95% confidence intervals overlap completely. Among the top 50, they are effectively indistinguishable. At the sample sizes typical of world-class athlete studies, 50 to 100 participants (Mitchell et al., 2024), a weak positive effect and a weak negative effect generate almost the same within-elite pattern.

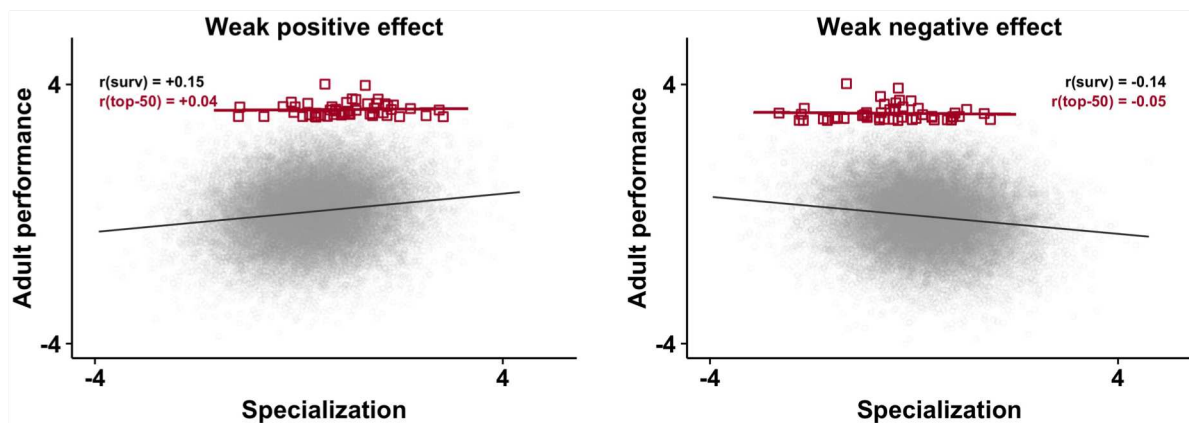
Panel B shows why this is the case. In the full survivor sample, the regression slope still separates the two scenarios: upward in Scenario B, downward in Scenario C. However, inside the elite subset, the fitted lines look essentially the same. Once attention is restricted to the selected tail, the information needed to distinguish the two developmental stories has disappeared.

Panel C illustrates the same phenomenon from a different angle. In the elite range, negative specialization-adult correlations are common under both scenarios. A study can easily recover the wrong sign when the true effect is weakly positive, but it can also fail to recover the negative sign when the true effect is weakly negative. Under both scenarios, roughly 35–45% of simulated cohorts recover the wrong sign. In this range, the sign of the within-elite association is not a reliable guide to the sign of the underlying effect. Only when the elite cutoff includes more than 5,000 adults do the two curves clearly separate. At the cutoffs where most studies of world-class performers operate, elite-only data can therefore make a weak positive effect look much like a negative one.

A Compression of the specialization signal



B Opposite effects, indistinguishable elite patterns



C The sign is unstable at small K

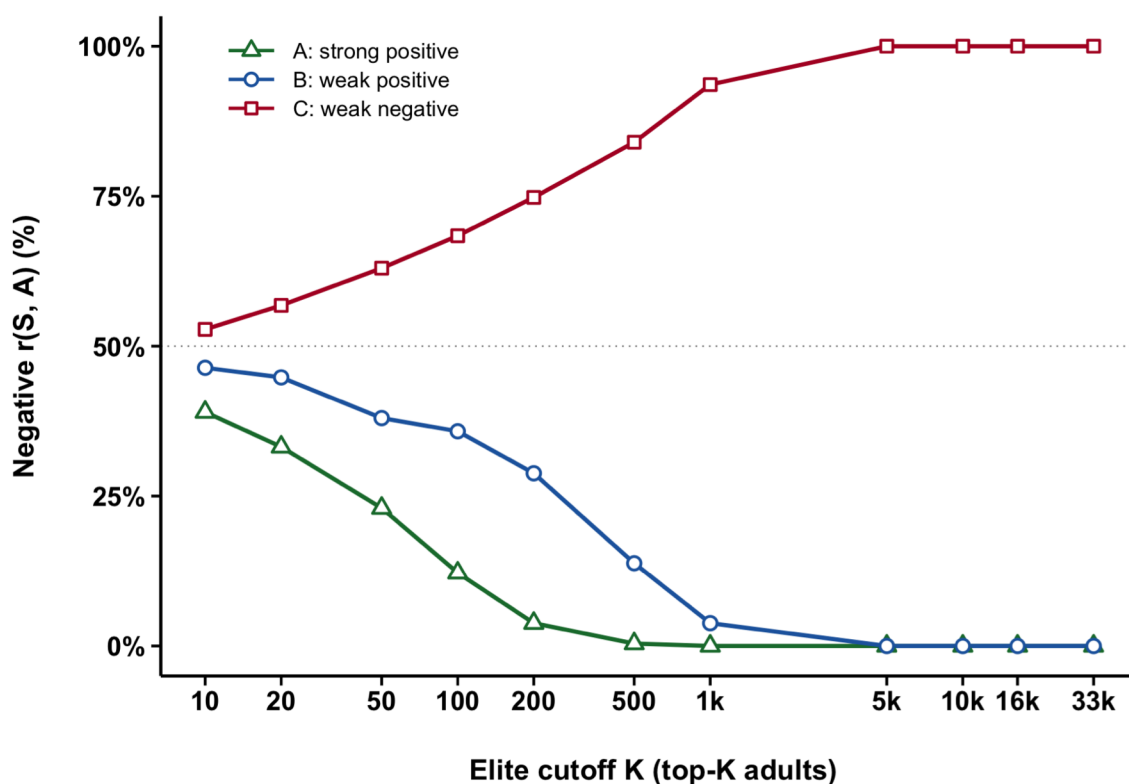


Figure 3. Selection compresses the specialization signal in the elite tail. (A) Correlation between specialization and adult performance ($r(S, A)$) as progressively more of the highest-performing adults are included, from the most selective group to all adult survivors. Squares at the far right show the corresponding population values; ribbons show 95% CIs across 500 simulated cohorts. (B) **Representative survivor-level and elite-only scatterplots for the weak-positive and weak-negative carry-over scenarios.** The full-sample slope (gray line) distinguishes the scenarios; the within-elite slope (red dashed) does not. (C) **Percentage of simulated cohorts in which the specialization-adult-performance correlation is negative, $r(S, A) < 0$, at each elite cutoff.** When the true effect is strongly positive, negative estimates become rare as more adults are included. By contrast, weak positive and weak negative effects are difficult to distinguish in the most selective groups because the sign is unstable in both cases.

The same logic applies beyond specialization. If practice explains substantial variance in broad samples but little variance among elites (Campitelli & Gobet, 2011; Ericsson et al., 1993; Hambrick et al., 2020), then that attenuation need not mean that practice matters

less at high levels. It may simply mean that selection has removed much of the variation that made the relation visible in the first place.

Repeating the simulation across a broad range of youth-to-adult relationships produced the same result (see SM, Section 3). Selection reduced both positive and negative population associations toward zero. Distinguishing them reliably required approximately 5,000 elite performers – around 100 times the sample size in most studies of world-class performers (e.g., Mitchell et al., 2024). Even adding a direct negative effect of specialization on adult performance did not allow elite-only data to distinguish the two explanations, because selection had already reduced the observed association to near zero (see SM, Section 3).

These simulations therefore show that selection is sufficient to produce the within-elite attenuation or reversal pattern under the developmental structures modeled here. They do not estimate how often selection accounts for such patterns in the published literature, nor do they rule out genuine developmental change.

Empirical Illustrations

The simulation shows that selection on adult performance can compress and even reverse the early-performance association among elites even when the underlying relationship is positive. The next question is whether the same pattern appears in longitudinal performance records.

We examine this in two domains with different measurement and developmental contexts. Chess offers a particularly clean case: performance is measured on a single continuous scale from childhood to adulthood (Elo, 1978; Vaci et al., 2019; Vaci & Bilalić, 2017), with no pronounced physical maturation component. Women's 1500m running provides a contrast: performance is physical, maturation is more relevant, and the survival

gate is harsher⁴. If the same threshold-dependent pattern appears in both, it is unlikely to be a peculiarity of a single domain.

We first examine the first gate: who remains observable as an adult (Figure 4)? In chess, the youth cohort included 43,078 players rated by age 14, of whom 16,421 (38%) still had a rating at age 20 or later. In women's 1500m, the youth cohort included 9,571 athletes with a recorded time before age 15, of whom only 688 (7%) recorded a time after age 20. Whether in chess or running, better youth performers were much more likely to remain visible in adulthood. Attrition was therefore related to youth performance. In both domains, those observed as adults were already a selected group before any elite cutoff was applied.

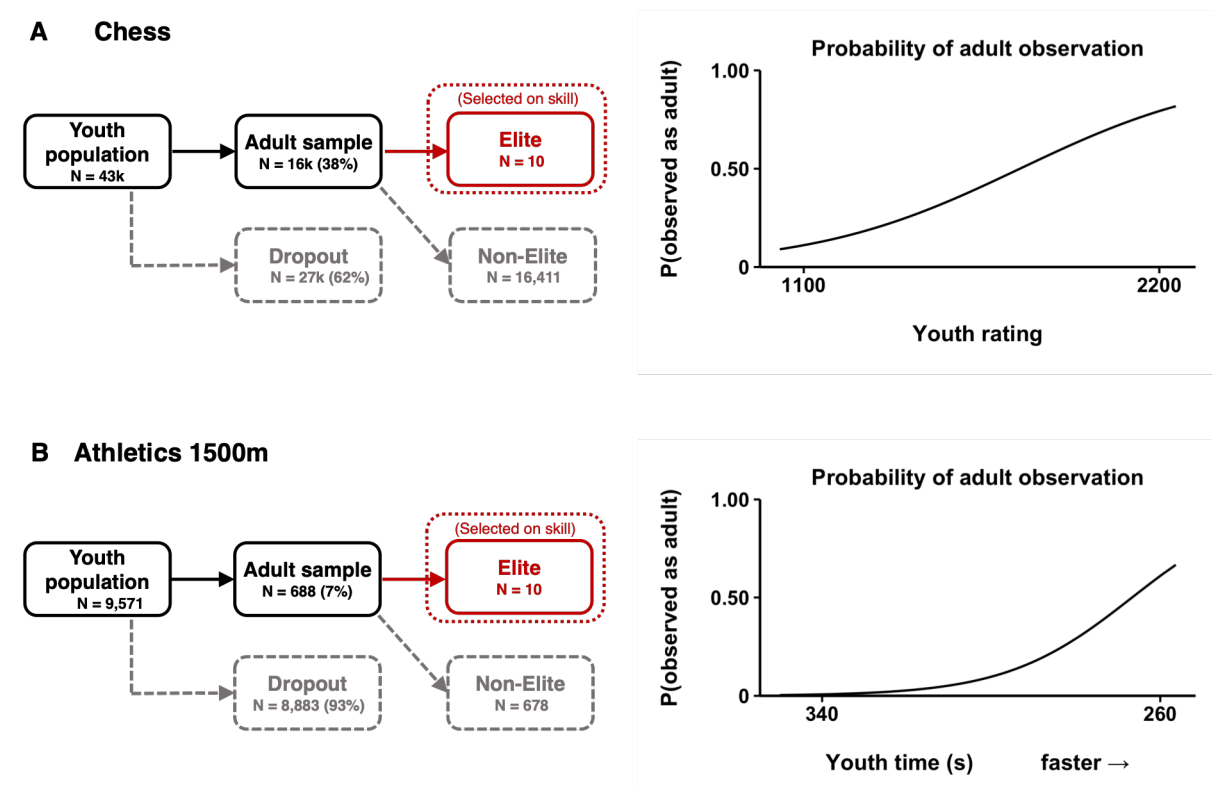


Figure 4. The first gate: better youth performers are more likely to remain observable as adults. (A) Chess. Left: sequential selection from the youth cohort (players rated by age 14) to the adult sample (players with a rating at age 20 or later) and then to an elite subset

⁴ This also means that the trajectory in athletics rests on more assumptions than the chess case, where a single rating scale runs from childhood to adulthood.

comprising those with the highest adult peak ratings. Right: predicted probability of adult observation as a function of youth rating. **(B) Women's 1500m.** Left: sequential selection from the youth cohort (athletes with a recorded time before age 15) to the adult sample (athletes with a recorded time after age 20) and then to an elite subset comprising those with the fastest adult times. Right: predicted probability of adult observation as a function of youth time (faster times toward the right). In both domains, stronger youth performance is associated with a higher probability of adult observation, showing that the adult sample is already selected before any elite cutoff is applied.

We then turn to the second gate: what happens once attention is restricted to the elite tail (Figure 5)? Among chess survivors, youth rating and adult peak rating were strongly positively related overall ($r = .82$). Yet, within the top-10 adults, the association reversed ($r = -.50$). In women's 1500m, youth and adult times were also positively related among survivors ($r = .35$), but again the association reversed within the top-10 ($r = -.13$). In both domains, the right-hand plots tell the same story. Among the most selective groups, the youth-adult association becomes weak or negative. As progressively more adults are included, it recovers toward the value observed among all adult survivors.

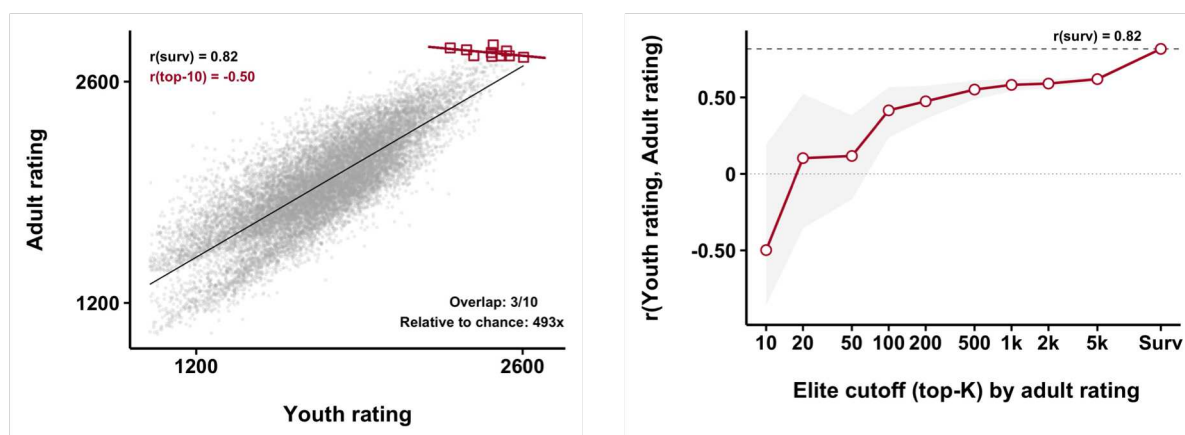
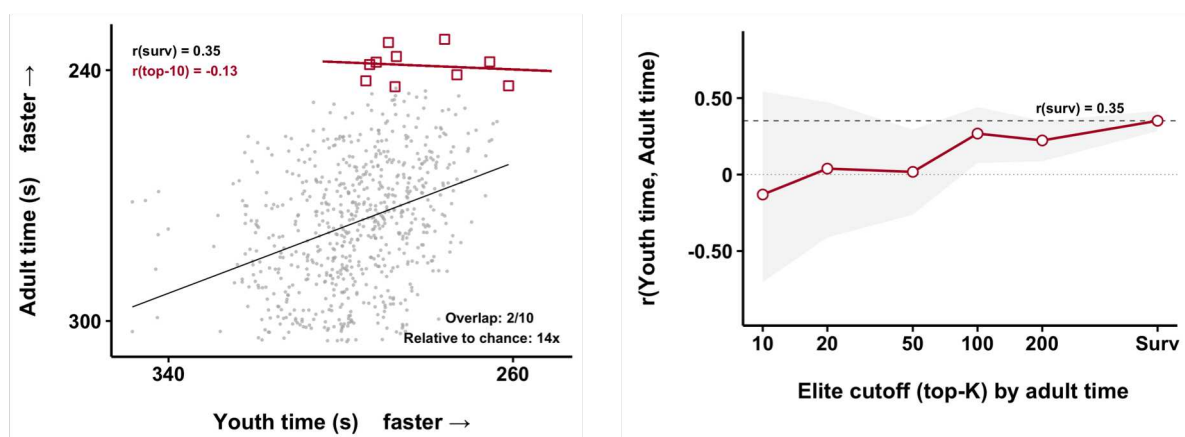
A Chess**B Women's 1500m**

Figure 5. The second gate: the youth–adult association changes across elite thresholds. (A) Chess. Left: youth rating versus adult peak rating among adult survivors (surv), with the top-10 adults highlighted. Right: correlation between youth and adult performance as progressively more of the highest-performing adults are included. (B) Women's 1500m. Left: youth time versus adult best time among adult survivors (axes reversed so faster performances appear toward the top right), with the top-10 adults highlighted. Right: correlation between youth and adult performance as progressively more of the highest-performing adults are included. In both domains, the youth–adult association is positive among adult survivors but becomes weak or negative in the most selective groups, then recovers as more adults are included. Its direction and strength therefore depend on how narrowly the elite group is defined. Gray ribbons show 95% confidence intervals (Fisher's z). Inserts (bottom right) summarize overlap between the top-10 youth and top-10 adults relative to chance (enrichment).

The two domains differ in strength, but not in form. In chess, the youth to adult relation remains strong after moderate selection and reverses only in the extreme tail. In

running, the relation is weaker to begin with and breaks down much earlier, with estimates spanning both positive and negative values across much of the elite range. This mirrors the pattern from the simulation: a strong relation survives until the most extreme cutoffs, a weaker one collapses sooner⁵.

A common response to such findings is to point to the low overlap between top juniors and top adults and to treat that as evidence that youth success does not predict adult excellence (Barth et al., 2022; Güllich et al., 2023; Güllich et al., 2025, but see Bezuglov et al., 2022). But low absolute overlap is not the right test. The relevant question is whether the observed overlap exceeds what chance alone would predict within the adult survivor pool. In chess, 3 of the top-10 juniors also appeared in the top-10 adults. That seems small in absolute terms, but chance predicts fewer than 0.01 such overlaps among 16,421 survivors ($10 \times 10 / 16,421$). The observed overlap is therefore about 493 times chance. In Women's 1500m, 2 of the top-10 juniors appeared in the top-10 adults, compared with an expected overlap of 0.15 among 688 survivors – about 14 times chance.

Even modest absolute overlap can be consistent with a strong positive relationship among all adult survivors when elite status is rare. In chess, for example, even with $r = .82$ among survivors, only 3 of the top-10 juniors reappear in the adult top-10. The issue is not that youth performance is irrelevant. Rather, once selection becomes extreme, even a positive developmental relationship leaves only limited overlap at the very top.

Principles for Design and Reporting

⁵ These analyses illustrate the empirical pattern but they do not establish what caused it. They show that the youth–adult association changes across elite thresholds in real longitudinal data, but not that selection caused the reversals reported elsewhere. A reversal among elites could arise because selection has obscured a positive relationship, but it could also reflect a genuinely weak or negative developmental effect. The elite data cannot distinguish these possibilities on their own.

We believe that the implications extend beyond a technical correction to how expertise research is designed, analyzed, reported, and interpreted. To translate these implications into practice, we propose four principles for the study of outstanding performance.

1. Decide what you want to learn. Outstanding-performance research serves two goals that are routinely conflated. One is prediction within an already-selected pathway: among those admitted to an academy, who progresses further? The other asks what early experiences increase the probability of later excellence among the broader population of aspirants? The study design should follow from the goal. The first can proceed with an elite sample. The second requires either a population-based cohort or explicit modeling of the selection that produced the sample (Loh & Ren, 2025).

2. Define the entry cohort and map the gates. When the research question concerns developmental pathways, studies should start from an identified entry cohort at a clear baseline, track who is lost at each stage, and report the numbers. If one starts from the adult sample, then potentially consequential earlier selection is no longer visible.

3. Show how results depend on selection. Figures 3 and 5 show that associations among elite performers can change sharply as the elite cutoff changes. Researchers should therefore show whether their results hold across different cutoffs and stages of selection, as well as whether overlap exceeds chance. In very small elite samples, apparent reversals may also reflect sampling variability.

4. Match the claim to the sample. A simple rule is to say exactly who was studied: "among those still competing at age 22," "within the top 50 by adult performance." Unless selection has been explicitly addressed, such findings should be read as descriptions of that selected group, not of how excellence develops more generally.

Designs and Analyses That Address Selection

Existing datasets often include only adult survivors, academy members, or elite performers. Some of the missing pathway can nevertheless be reconstructed. When earlier measurements predict who remains observable as an adult, inverse-probability weighting gives more weight to observed individuals who resemble those most likely to have left. This moves the sample closer to the original cohort (Chatton & Rohrer, 2024; Hernán & Robins, 2020). When dropout may also depend on factors that were not measured, sample-selection models can help, but only when their assumptions can be justified (Heckman, 1979). If selection has mainly narrowed the range of performance, range-restriction corrections can estimate how much the association may have been reduced (Sackett & Yang, 2000; Thorndike, 1949). Negative controls, that is, predictors or outcomes that should not be causally related, can provide a further check on whether bias is a plausible explanation (Lipsitch et al., 2010). Further details are provided in the SM (Section 5).

Where feasible, prospective designs can go further by altering selection itself. Randomizing admission among borderline candidates (e.g., ethical lottery) weakens the link between admission and the predictors of interest among otherwise eligible candidates. It also retains variation that selection would otherwise remove, which in turn allows estimates more directly relevant to admission policy (Abdulkadiroğlu et al., 2011; Angrist et al., 1996; ten Cate, 2021).

Researchers can reconstruct developmental pathways only if the relevant information has been recorded along the way. Federations, academies, and other organizations can support this work by documenting who was eligible at the outset, who passed through each successive gate, and why others left. Without these records, the role of access, support, injury, financial constraints, and institutional decisions in determining who

remains is lost from view. When researchers observe only those who remain into adulthood, unequal opportunities may therefore appear as different developmental routes to the same outcome (Gamoran, 2002; Kratz & Patzina, 2020).

Conclusion

Within-elite reversals in deliberate practice, early specialization, or youth performance do not by themselves show that developmental processes change at the highest levels. A two-gate simulation, which includes survival followed by elite thresholding, can produce similar patterns without requiring a negative developmental mechanism – a pattern mirrored in the empirical analyses from chess and athletics. This does not mean that developmental explanations are wrong, or that selection and developmental change cannot operate together. It does, however, mean that, at the cutoffs and sample sizes these studies use, elite-only data cannot reliably distinguish the proposed developmental explanations from selection-induced patterns.

The lesson here is not that elite samples are uninformative, but instead that they answer a different question from the one they are often asked to answer. When expertise is the predictor, expert–novice comparisons can reveal how skilled perception, memory, attention, and decision-making are organized (e.g., Costantino et al., in press). When expertise is the outcome, however, samples restricted to those who have already succeeded cannot by themselves reveal the developmental causes of that success. In that setting, the sample has already been shaped by participation, retention, opportunity, and performance gates.

Outstanding performance will always be rare, and multiple routes will often lead to the same exceptional outcome. Selection is therefore not a nuisance that better methods can simply wash away. It is built into the structure of any domain where many begin and

SELECTION BIAS IN EXPERTISE RESEARCH

few arrive at the top. If we want to learn from excellence, we must make these selection processes explicit.

References

- Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, 126(2), 699–748.
<https://doi.org/10.1093/qje/qjr017>
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.1080/01621459.1996.10476902>
- Barth, M., Güllich, A., Macnamara, B. N., & Hambrick, D. Z. (2022). Predictors of junior versus senior elite performance are opposite: A systematic review and meta-analysis of participation patterns. *Sports Medicine*, 52(6), 1399–1416.
<https://doi.org/10.1007/s40279-021-01625-4>
- Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3), 47–53. <https://doi.org/10.2307/3002000>
- Bezuglov, E., Emanov, A., Waśkiewicz, Z., Semeniuk, N., Butovsky, M., Shoshorina, M., Baranova, D., Volodina, K., & Morgans, R. (2022). Successful young athletes have low probability of being ranked among the best senior athletes, but this is higher when compared to their less successful peers. *Frontiers in Psychology*, 13, 869637.
<https://doi.org/10.3389/fpsyg.2022.869637>
- Bilalić, M. (2017). *The neuroscience of expertise*. Cambridge University Press.

- Bilalić, M. (2018). The double take of expertise: Neural expansion is associated with outstanding performance. *Current Directions in Psychological Science*, 27(6), 462-469. <https://doi.org/10.1177/0963721418793133>
- Bilalić, M., McLeod, P., & Gobet, F. (2007). Does chess need intelligence?—A study with young chess players. *Intelligence*, 35(5), 457-470. <https://doi.org/10.1016/j.intell.2006.09.005>
- Binet, A. (1894). *Psychologie des grands calculateurs et joueurs d'échecs*. Hachette.
- Campitelli, G., & Gobet, F. (2011). Deliberate practice: Necessary but not sufficient. *Current Directions in Psychological Science*, 20(5), 280–285. <https://doi.org/10.1177/0963721411421922>
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81. [https://doi.org/10.1016/0010-0285\(73\)90004-2](https://doi.org/10.1016/0010-0285(73)90004-2)
- Chatton, A., & Rohrer, J. M. (2024). The causal cookbook: Recipes for propensity scores, g-computation, and doubly robust standardization. *Advances in Methods and Practices in Psychological Science*, 7(1), 25152459241236149. <https://doi.org/10.1177/25152459241236149>
- Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2), 417–420. <https://doi.org/10.1093/ije/dyp334>

Costantino, A. I., Platonov, A., Fontana Vieira, F., Van Hove, E., Bilalić, M., & Op de Beeck, H.

(in press). Low-dimensional and optimised representations of high-level information in the expert brain. *Nature Communications*. <https://doi.org/10.1038/s41467-026-74566-z>

de Groot, A. D. (1978). *Thought and choice in chess*. Mouton. (Original work published 1946)

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Publishing.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1), 31–53.
<https://doi.org/10.1146/annurev-soc-071913-043455>

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406.
<https://doi.org/10.1037/0033-295X.100.3.363>

Ericsson, K. A., & Harwell, K. W. (2019). Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance: Why the original definition matters and recommendations for future research. *Frontiers in Psychology*, 10, 2396. <https://doi.org/10.3389/fpsyg.2019.02396>

Ford, P., De Ste Croix, M., Lloyd, R., Meyers, R., Moosavi, M., Oliver, J., Till, K., & Williams, C. (2011). The Long-Term Athlete Development model: Physiological evidence and application. *Journal of Sports Sciences*, 29(4), 389–402.
<https://doi.org/10.1080/02640414.2010.536849>

Gamoran, A. (2002). *Standards, inequality & ability grouping in schools* (Briefing No. 25).

Centre for Educational Sociology, University of Edinburgh.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001).

Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 236-243.

[https://doi.org/10.1016/S1364-6613\(00\)01662-4](https://doi.org/10.1016/S1364-6613(00)01662-4)

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology*, 31(1), 1-40.

<https://doi.org/10.1006/cogp.1996.0011>

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243-1255. <https://doi.org/10.1177/1745691620921521>

Güllich, A., & Barth, M. (2024). Effects of early talent promotion on junior and senior performance: A systematic review and meta-analysis. *Sports Medicine*, 54(3), 697-710. <https://doi.org/10.1007/s40279-023-01957-3>

Güllich, A., Barth, M., Hambrick, D. Z., & Macnamara, B. N. (2025). Recent discoveries on the acquisition of the highest levels of human performance. *Science*, 390, eadt7790. <https://doi.org/10.1126/science.adt7790>

Güllich, A., Barth, M., Macnamara, B. N., & Hambrick, D. Z. (2023). Quantifying the extent to which successful juniors and successful seniors are two disparate populations: A systematic review and synthesis of findings. *Sports Medicine*, 53(6), 1201-1217. <https://doi.org/10.1007/s40279-023-01840-1>

Güllich, A., Macnamara, B. N., & Hambrick, D. Z. (2022). What makes a champion? Early multidisciplinary practice, not early specialization, predicts world-class performance. *Perspectives on Psychological Science*, 17(1), 6–29.

<https://doi.org/10.1177/1745691620974772>

Hambrick, D. Z., Macnamara, B. N., & Oswald, F. L. (2020). Is the deliberate practice view defensible? A review of evidence and discussion of issues. *Frontiers in Psychology*, 11, 1134. <https://doi.org/10.3389/fpsyg.2020.01134>

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert? *Intelligence*, 45, 34–45. <https://doi.org/10.1016/j.intell.2013.04.001>

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. <https://doi.org/10.2307/1912352>

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5), 615–625. <https://doi.org/10.1097/01.ede.0000135174.63482.43>

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Howe, C. J., Cole, S. R., Lau, B., Napravnik, S., & Eron, J. J. (2016). Selection bias due to loss to follow up in cohort studies. *Epidemiology*, 27(1), 91–97. <https://doi.org/10.1097/EDE.0000000000000409>

Kratz, F., & Patzina, A. (2020). Endogenous selection bias and cumulative inequality over the life course: Evidence from educational inequality in subjective well-being. *European Sociological Review*, 36(3), 333–350. <https://doi.org/10.1093/esr/jcaa003>

Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, 6(2), 25152459231156085. <https://doi.org/10.1177/25152459231156085>

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3), 383–388. <https://doi.org/10.1097/EDE.0b013e3181d61eeb>

Loh, W. W., & Ren, D. (2025). From experiments to policy insights: Generalizing causal effects from study samples to target populations. *Advances in Methods and Practices in Psychological Science*, 8(3), 25152459251359623. <https://doi.org/10.1177/25152459251359623>

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science*, 25(8), 1608–1618. <https://doi.org/10.1177/0956797614535810>

Macnamara, B. N., Moreau, D., & Hambrick, D. Z. (2016). The relationship between deliberate practice and performance in sports: A meta-analysis. *Perspectives on Psychological Science*, 11(3), 333–350. <https://doi.org/10.1177/1745691616635591>

McCarthy, J. (1990). Chess as the drosophila of AI. In T. A. Marsland & J.

Schaeffer (Eds.), *Computers, chess, and cognition* (pp. 227–237).

Springer. https://doi.org/10.1007/978-1-4613-9080-0_14

Mitchell, L., Ratcliff, J., Burke, L. M., & Forsyth, A. (2024). Engaging athletes as research participants: A document analysis of published sport science literature. *European Journal of Sport Science*, 24(10), 1442–1451. <https://doi.org/10.1002/ejsc.12198>

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1), 226–235.

<https://doi.org/10.1093/ije/dyx206>

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Rees, T., Hardy, L., Güllich, A., Abernethy, B., Côté, J., Woodman, T., Montgomery, H., Laing, S., & Warr, C. (2016). The Great British Medalists Project: A review of current knowledge on the development of the world's best sporting talent. *Sports Medicine*, 46(6), 1041–1058. <https://doi.org/10.1007/s40279-016-0476-2>

Ren, D., & Loh, W. W. (2024). Advancing group-based disparities research and beyond: A cautionary note on selection bias. *Advances in Methods and Practices in Psychological Science*, 7(3), 25152459241260256. <https://doi.org/10.1177/25152459241260256>

Richardson, T. G., Davey Smith, G., & Munafò, M. R. (2019). Conditioning on a collider may induce spurious associations: Do the results of Gale et al. (2017) support a health-

protective effect of neuroticism in population subgroups? *Psychological Science*, 30(4), 629–632. <https://doi.org/10.1177/0956797618774532>

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>

Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: Causality and the within-/between-persons distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, 6(1), 25152459221140842. <https://doi.org/10.1177/25152459221140842>

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1), 112–118. <https://doi.org/10.1037/0021-9010.85.1.112>

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist*, 61(4), 394–403.

Southwick, D. A., Harwell, K. W., Wright, G., Olsen, J. A., & Ogles, B. M. (in press). Not All Practice Is Created Equal: Longitudinal Evidence From Over 40,000 Chess Players. *Psychological Science*, <https://doi.org/10.1177/09567976261452568>

ten Cate, O. (2021). Rationales for a lottery among the qualified to select medical trainees: Decades of Dutch experience. *Journal of Graduate Medical Education*, 13(5), 612–615. <https://doi.org/10.4300/JGME-D-21-00789.1>

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. Wiley.

- Till, K., & Baker, J. (2020). Challenges and [possible] solutions to optimizing talent identification and development in sport. *Frontiers in Psychology*, 11, 664.
<https://doi.org/10.3389/fpsyg.2020.00664>
- Ullén, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. *Psychological Bulletin*, 142(4), 427–446. <https://doi.org/10.1037/bul0000033>
- Vaci, N., & Bilalić, M. (2017). Chess databases as a research vehicle in psychology: Modeling large data. *Behavior Research Methods*, 49(4), 1227–1240.
<https://doi.org/10.3758/s13428-016-0782-5>
- Vaci, N., Edelsbrunner, P., Stern, E., Neubauer, A., Bilalić, M., & Grabner, R. H. (2019). The joint influence of intelligence and practice on skill development throughout the life span. *Proceedings of the National Academy of Sciences*, 116(37), 18363–18369.
<https://doi.org/10.1073/pnas.1819086116>
- Vaci, N., Gula, B., & Bilalić, M. (2014). Restricting range restricts conclusions. *Frontiers in Psychology*, 5, 569. <https://doi.org/10.3389/fpsyg.2014.00569>
- Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, 5(2), 25152459221095823. <https://doi.org/10.1177/25152459221095823>

Acknowledgments

We thank Robert Gaschler for comments on a previous draft of the manuscript. We also acknowledge the use of Claude (Anthropic) and ChatGPT (OpenAI) for assistance with manuscript editing and simulation code development. All output was reviewed and revised by the authors, who take full responsibility for the content.

Author Contributions

Merim Bilalić: Conceptualization, Methodology, Formal analysis, Data curation, Software, Visualization, Writing – original draft, Project administration

Nemanja Vaci: Conceptualization, Methodology, Writing – review & editing

Johannes Hönekopp: Conceptualization, Methodology, Writing – review & editing

Thomas Pollet: Conceptualization, Methodology, Writing – review & editing

Funding

N/A

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship or the publication of this article.