A Statistical Spectroscope for Fair Play and Growing Minds

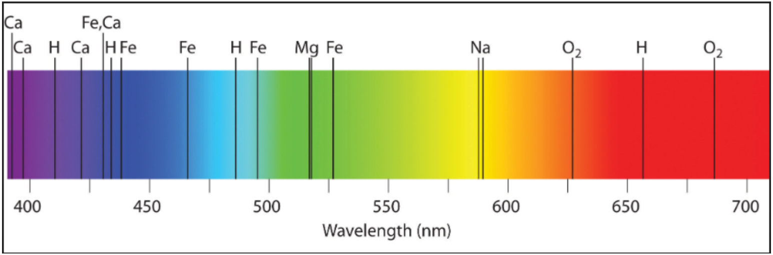
Smart Moves Summit 2025

Kenneth W. Regan[1]
University at Buffalo (SUNY)

5 August, 2025

[1] With grateful acknowledgment to co-authors Guy L. Haworth and Tamal Biswas, students in my graduate seminars, and UB's Center for Computational Research (CCR)

What G.L. Haworth meant by “Skilloscopy”



- Isolate and zoom in on **elements of chess skill** at all levels.
- *Assertion*: these follow from simpler *laws of (human!) cognition*.
- *Assertion*: **Human perception of value** is central.
- ⟶ **econometric** modeling of consumer value and preferences.
- Studying *perception* requires an **objective** benchmark of value.
- Powerful chess programs give that in chess. 🔥 ⟷ 🔥

Some General Themes

- Simple elements **Strategy** and **Tactics** take us far.
- **Depth of Thinking** should be next.
- Do weaker players **prefer** weaker moves?
- Or are they more easily **distracked?**
- **Logistic Curves Are Everywhere.**
- So is **Procrastination**.
- How shall we handle the element of **Difficulty**?
- **Recognition** "Versus" **Thinking**.
    - See the 2007 National Geographic documentary "My Brilliant Brain" with Susan Polgar (crux here).
    - We will try to glean comparable insight from numerical analytics.

A Predictive Analytic Model

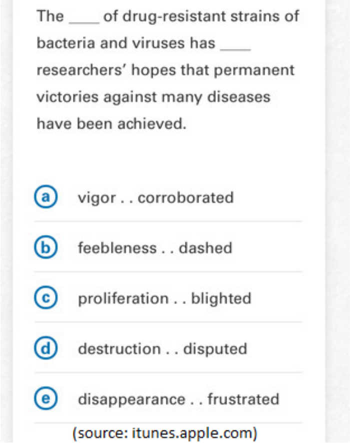
Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \dots, m_j, \dots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

In a **utility-based** model, each m_i has a utility or cost u_i.
Main risk/reward quantity then becomes $E = \sum_i p_i u_i$.

- **Insurance**: m_i are risk factors; costs u_i need not influence p_i.
- **Chess**: m_i are legal moves; u_i are engine values and influence p_i.
- **Multiple-choice tests**: m_i are possible answers to a test question, u_i = gain/loss for right/wrong answer.

Chess and Tests—With Partial Credits (Or LLMs?)

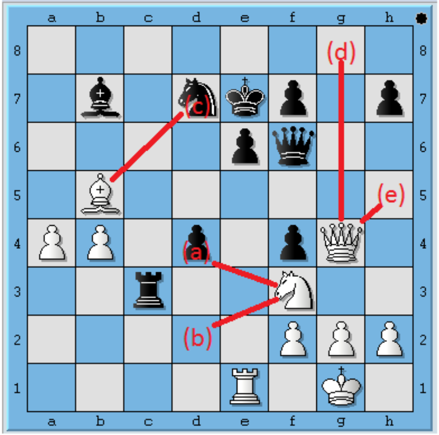


The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

- (a) vigor . . corroborated
- (b) feebleness . . dashed
- (c) proliferation . . blighted
- (d) destruction . . disputed
- (e) disappearance . . frustrated

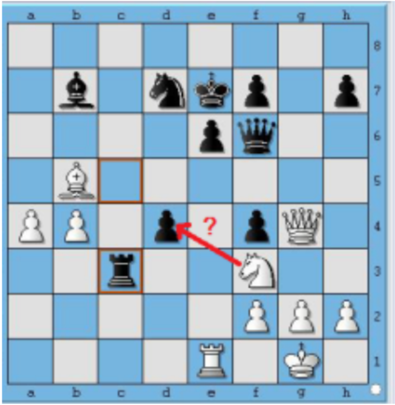
(source: itunes.apple.com)

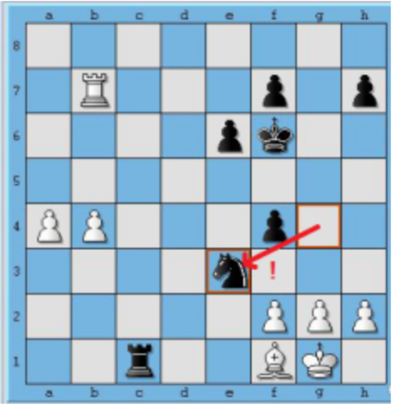
=



Here (b,c) are **equal-optimal** choices, (a) is bad, but (d) and (e) are reasonable—worth part credit.

Move Utilities Example (Kramnik-Anand, 2008)





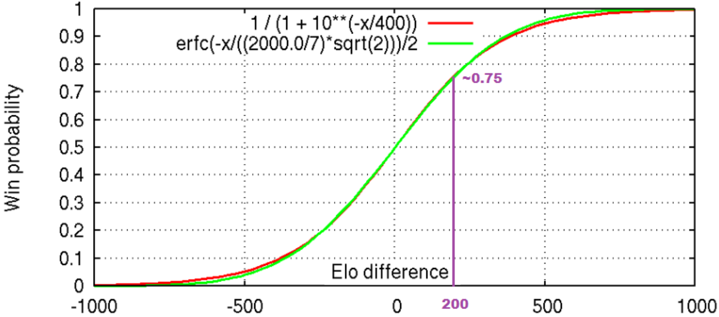
Depths...

Values by Stockfish 6

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

Aptitude—Via Elo Grades (calculator)

- Named for **Arpad Elo**, number R_P rates skill of player P.
- E.g. **1000** = bright beginner, **1600** = good club player, **2200** = master, **2800** = world championship caliber.
- Computer **engines** are far higher, e.g.: **Stockfish 16 = 3544, Torch 1.0 = 3531, Komodo Dragon 3.3 = 3529.**
- Expectation given by rating *difference* via this logistic curve:



Main Parameters and Inputs

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment. *Like Anatoly Karpov.*
- c for "consistency" in tactical minefields. *Like Mikhail Tal.*
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 with both players near the same Elo marker 1025, 1050, . . . , 2775, 2800, 2825.

Being retrained on new FIDE range **1400**. . . 2825, **from 1/1/25 on**.

- Given an Elo rating R, "central slice" gives corresponding s_R, c_R, h_R.
- Only other input is the grid of move utilities $u_{i,d}$ at various depths d of search, further **scaled** to make (perceived) values v_i (and ρ_i).
- Then $\delta_i = v_1 - v_i$ is difference to best move.
- Other than these, **my model knows nothing about chess.**

One Wonky Slide: Log-Linear Versus Loglog-Linear

The generic **log-linear** model puts

$$\log\left(\frac{1}{p_i}\right) = \alpha + \beta u_i, \quad \text{or equivalently,} \quad \log\left(\frac{1}{p_i}\right) - \log\left(\frac{1}{p_1}\right) = \beta\delta_i$$

- Solved by **softmax** giving $p_i = p_1 \cdot \exp(-\beta u_i)$.
- Each p_i is represented as a **multiple** of the top probability p_1.
- Ubiquitous in AI—but does not work for chess.

The **loglog-linear** model puts $\log\log(\frac{1}{p_i}) - \log\log(\frac{1}{p_1}) = \beta\delta_i$, i.e.:

$$\frac{\log(1/p_i)}{\log(1/p_1)} = \exp(\beta\delta_i).$$

- Gives $p_i = p_1^{\exp(\beta\delta_i)}$.
- So p_i are represented as **powers** of the best-move probability p_1.
- In place of $\beta\delta_i$, I really have $\left(\frac{\delta_i - h\rho_i}{s}\right)^c$, with h tightly clamped.

How The Model Operates

- Take s, c, h from a player's rating (or wider skill profile).
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, **1,000**p_i times.
- **Roll the die** to give confidence intervals that go with the p_i.
- (Correct after-the-fact for chess decisions not being independent.)

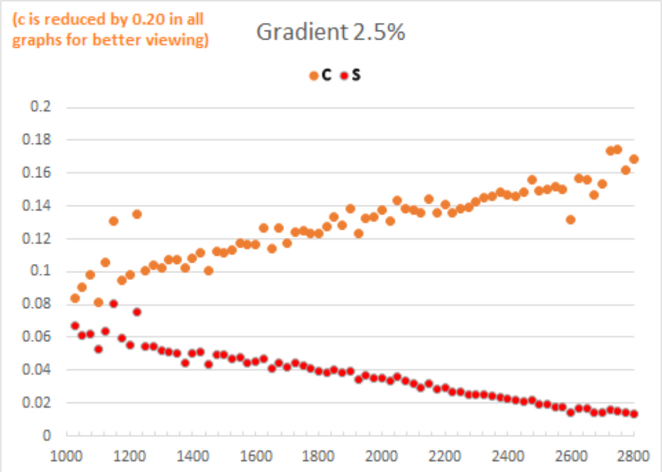
Main Outputs:

- **Statistical z-scores** for various (*actual*−*projected*) quantities:
  - **T1-match**: Agreement with the move listed first by the computer.
  - **EV-match**: Includes moves of equal-optimal value not listed first.
  - **ASD**: Average *scaled* difference in value from inferior moves.
- An **Intrinsic Performance Rating** (**IPR**) for the set of games.

Fit s, c, h by making T1,EV,ASD be **unbiased estimators** on the training sets, which are stratified by Elo ratings.

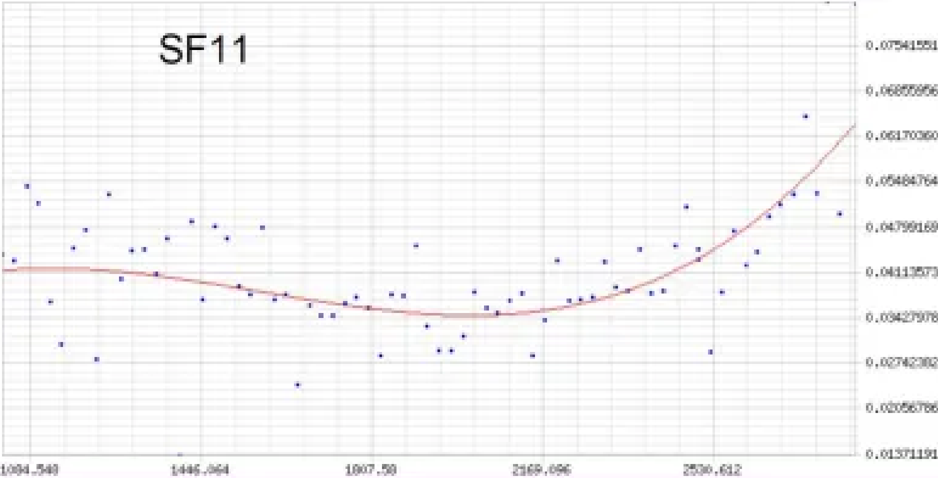
Parameters To Elo

Created in 2019. Note “noise” especially below 1250 and above 2575.



Predictive Accuracy (similar for other engines)

Error range $[p_i^{1-\epsilon} \dots p_i^{1+\epsilon}]$ for ϵ depending on rating as follows:



Good up to 3100 or so. Not bad for a **0.000000003B** parameter model.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per Stockfish 11): $s =$ **0.016,** $c =$ **0.307**.
- Tal (per Stockfish 11): $s =$ **0.026,** $c =$ **0.365**.
- Lower s is better—so Karpov was more "Karpovian."
- Higher c is better—so my model with Tal's parameters would make fewer large mistakes.

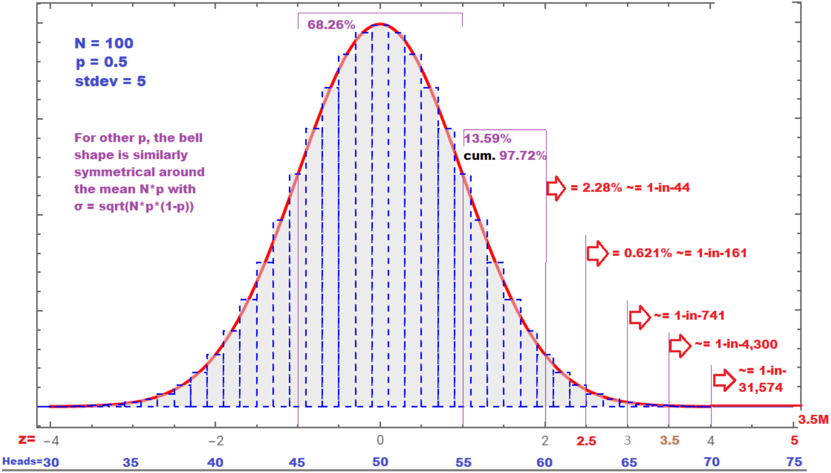
Are these grainy parameters enough to mimic human tendencies?

- IPRs: Karpov **2625 +- 155**, Tal **2730 +- 185**.
- Whole tourney IPR is (only!) **2575 +- 50** ($s = 0.041$, $c = 0.385$).
- Average Elo of players, **2621**, is within error bars. Surprise is that the IPR is not near 2700s range. Today's elite regularly hit 2800+.

Z-Scores

- A **z-score** measuresf performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called “sigmas” (σ).
- Correspond to statements of odds-against (**but see next slides**):
- “Six Sigma” (6σ) means about 500,000,000–1 odds;
- 5σ = 3,000,000–1;
- 4.75σ = 1,000,000–1;
- 4.5σ = 300,000–1;
- 4σ = 32,000–1;
- 3σ = 740–1;
- 2σ = 43–1 (civil minimum standard, polling “margin of error”).

Bell Curve and Tails



Blue = binomial 100 scale of the **screening stage**. WSTC examples.

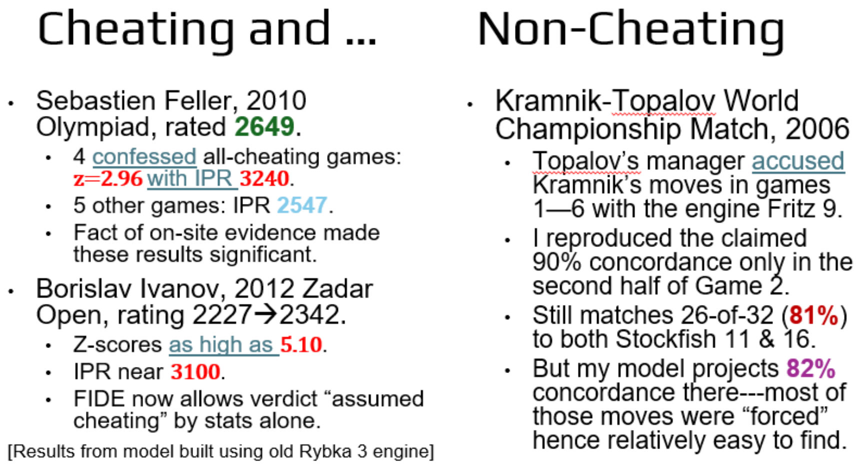
Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month? year?

Are these considerations orthogonal, or do they align?

Over large datasets from (presumably) non-cheating players, the **Central Limit Theorem** "kicks in" well: the z-scores conform to the bell curve.

Some Example Cases (old ones on-purpose...)



Cheating and ...

- Sebastien Feller, 2010 Olympiad, rated **2649**.
  - 4 confessed all-cheating games: **z=2.96** with IPR **3240**.
  - 5 other games: IPR **2547**.
  - Fact of on-site evidence made these results significant.
- Borislav Ivanov, 2012 Zadar Open, rating 2227→2342.
  - Z-scores as high as **5.10**.
  - IPR near **3100**.
  - FIDE now allows verdict “assumed cheating” by stats alone.

[Results from model built using old Rybka 3 engine]

Non-Cheating

- Kramnik-Topalov World Championship Match, 2006
  - Topalov’s manager accused Kramnik’s moves in games 1—6 with the engine Fritz 9.
  - I reproduced the claimed 90% concordance only in the second half of Game 2.
  - Still matches 26-of-32 (**81%**) to both Stockfish 11 & 16.
  - But my model projects **82%** concordance there---most of those moves were “forced” hence relatively easy to find.

Cognitive Studies and Chess Research

In **general cognitive research**, many results come from studies that

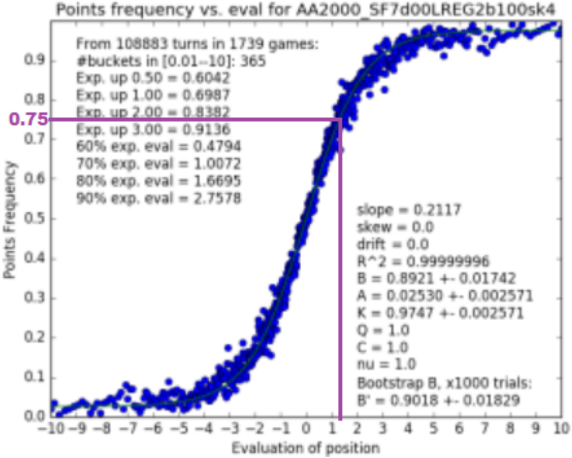
1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...
4. ...with unclear metrics and alignment of personal vs. test goals...,
5. ...and where reproducibility is doubtful and arduous.

Per my Daniel Kahneman obit, we should trade 1 against wealth of 2,3,4,5: lots of players and games, real competition, clear goals and metrics, reproducible, and conducive to abundant falsifiable predictions.

Here our subject **is** chess, so no problem!

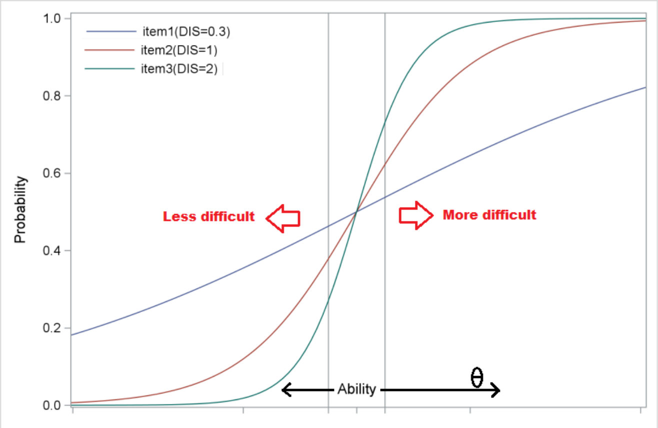
Let's consider elements of **difficulty** and **time pressure**.

Position Value ⟷ Expectation (2000 vs. 2000)



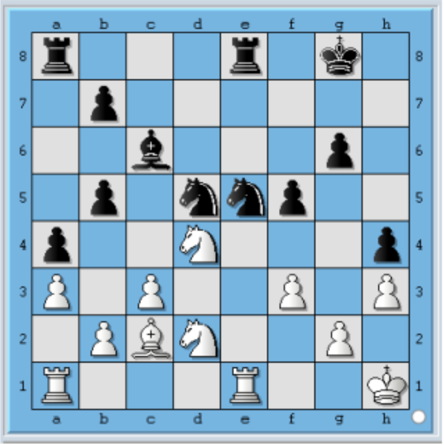
- Similar **0.75** expectation when up 1.30 vs. equal-rated player.
- Does **difficulty** equate to *expected value loss* (which I call **hazard**)?

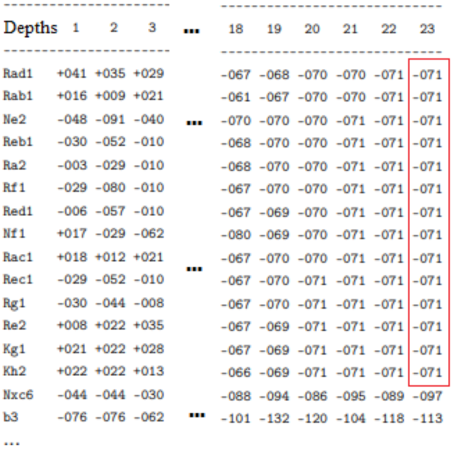
Item-Response Theory (IRT source)



- Horizontal axis governs **difficulty** in relation to $\theta =$ **ability**.
- Slope at $y = 0.5$ *correctness rate* is the **discrimination** factor.
- Difficulty ≈ expected (loss of) grading points. Recall $E = \sum_i p_i u_i$.

But see: Niemann-Shankland, USA Ch. 2023





Depths	1	2	3	...	18	19	20	21	22	23
Rad1	+041	+035	+029		-067	-068	-070	-070	-071	-071
Rab1	+016	+009	+021		-061	-067	-070	-070	-071	-071
Ne2	-048	-091	-040	...	-070	-070	-070	-071	-071	-071
Reb1	-030	-052	-010		-068	-070	-070	-071	-071	-071
Ra2	-003	-029	-010		-068	-070	-070	-071	-071	-071
Rf1	-029	-080	-010		-067	-070	-070	-071	-071	-071
Red1	-006	-057	-010		-067	-069	-070	-071	-071	-071
Nf1	+017	-029	-062		-080	-069	-070	-071	-071	-071
Rac1	+018	+012	+021	...	-067	-070	-070	-071	-071	-071
Rec1	-029	-052	-010		-067	-070	-071	-071	-071	-071
Rg1	-030	-044	-008		-067	-070	-071	-071	-071	-071
Re2	+008	+022	+035		-067	-069	-071	-071	-071	-071
Kg1	+021	+022	+028		-067	-069	-071	-071	-071	-071
Kh2	+022	+022	+013		-066	-069	-071	-071	-071	-071
Nxc6	-044	-044	-030		-088	-094	-086	-095	-089	-097
b3	-076	-076	-062	...	-101	-132	-120	-104	-118	-113
...										

Low-hazard because crisis is far off, but difficult in real chess terms. Low E, but high **entropy** from many (yucky) choices. (Niemann lost.)

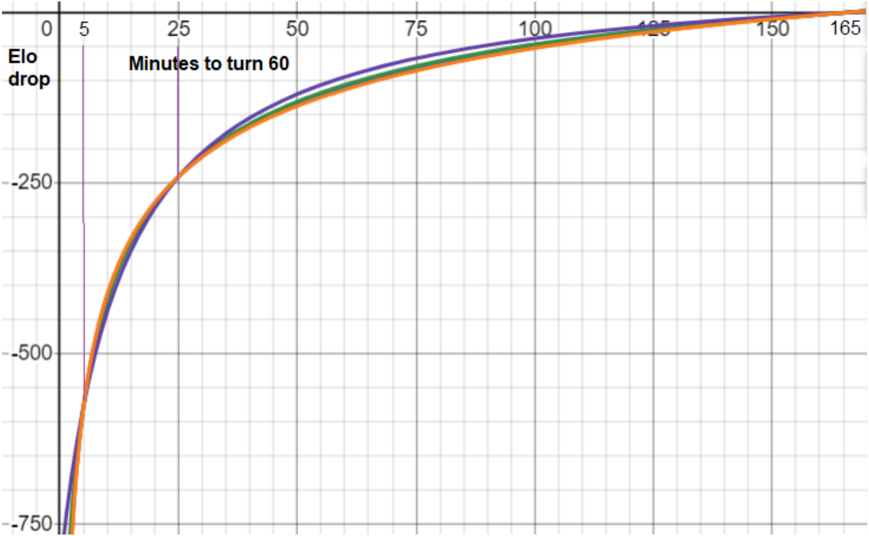
Aspects of Difficulty (Besides Hazard)

1. **Needing deep cogitation to find best move or avoid a trap**. *Expressly modeled—e.g. to project the trap for Kramnik.*
2. **Being at a disadvantage.** *Applies to chess, not so much examinations. Model performs fine.*
3. **Humans perform poorly.** *Basic with* **repeatable** *test questions.* Repeatable chess positions, however, are *opening book knowledge*.
4. **Humans take a long time to answer.**
5. **Question is inherently complex or taxing.**
    - How to measure this internally?
    - Sunde, Zegners, and Strittmatter [SZS, Jan. 2022] propose counting the time (i.e., number of position nodes) needed by chess engine to complete analysis to depth (say) 24.
    - Carow and Witzig [CW, Feb. 2024] consider all the above, but strive for human-chess based measures.

Time Budget and Effect on Quality

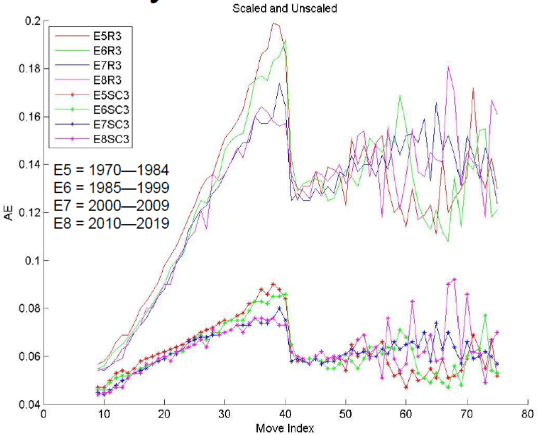
- **FIDE Standard Time Control**: 90 minutes to turn 40, then 30 minutes more, with 30-second *increment* after every move. Allows **150** minutes to turn 60.
- "Standard" control must allow at least **120** minutes to turn 60.
- Some elite events allow **180**, **195**, even **210** minutes (to turn 60).
- **Rapid** means any time giving under **60** minutes and at least **10**. Common is 15 min. plus 10-second increment, giving **25** to turn 60. Time control **2700+10** in use here can be called "semi-rapid."
- **Blitz** means under 10 minutes, most common is 3 minutes + 2-second increment, which gives **5** minutes to turn 60 — and so approximates old-school 5-minute chess on analog clocks.
- For **900+10** Rapid, I measure**d in 2015** a **240** reduction in quality.
- For **180+2** Blitz, **575** lower. (Error bars for both are about ±25.)

Time-Quality Curves (whole graph)



Time Usage, Procrastination, and Centipawn Loss

Error By Move Number in Games



Effect of time pressure approaching Move 40 is clear.

Moves 17—32 bridge between opening theory and worst of Zeitnot.

Mainly tournaments with lump of extra time after turn 40 up thru 2015.
Can imagine curve without a turn-40 sum (even with increment).
(How) Can we teach kids to use time more like the young Vishy Anand?

Predicated on Time Spent For a Move

IPRs of players rated **2000** to **2200** at the 2024 World Sr. Team Ch. in:

- Positions on which they spent at most **30 seconds** on the move: **2860 +- 75**.
- At most **10** seconds: **3235 +- 90**.
- Starting at turn 16 rather than 9: **3220 +- 100**.
- At most **5** seconds (sample size 605): **3230 +- 160**.

What gives here? How about moves with long thinks—?

- Positions with 5–10 minutes consumed: **1460 +- 85**.
- Using 10–15 minutes (705 positions): **1235 +- 170**.
- Using $\geq$ 15 minutes (371 positions): **1410 +- 205**.
- **"Thinking Is Bad For You."** (At least it's a bad sign...)
- Vivid reproduction of [SZS 2022] (and also Anderson et al., 2016 thru now for online blitz). "Think before you act...but not too long."

Instead of Seniors, Let's try 8-Year-Olds!

After 3 rounds of the **2024 World Cadets Championships** in separate Open and Girls' sections of ages **U08**, **U10**, and **U12**.

- The two **U08** sections combined have average rating 1596.
- I measure IPR as **1525 +- 45**. (10,913 positions total)
- In EWN mode, **1490 +- 65**.
- Positions on which they spent at most **30 seconds** on the move: **2170 +- 125** (2,996 pos.)
- At most **10** seconds: **2860 +- 245** (632 positions)
- At most **5** seconds (sample size 151): **2935 +- 555**.

How about when little kids think longer?

- Positions with 5–10 minutes consumed (729 pos.): **650 +- 235**.
- Using 10–15 minutes (168 positions): **465 +- 565**.
- Using $\geq$ 15 minutes (104 positions): **700 +- 505**.
- **"Thinking Is Bad For Kids Too."** (Reproduces at WTSC now.)

Hazard Vs. Time—and Time Left

Switching to Komodo 13.3 in place of Stockfish 11 as analyzing engine:

- Overall IPR of Elo 2000-to-2200 players: **2175 +- 35**.
- Average thinking time over all moves (turns 9–60): **181 seconds**.
- IPR on turns of $\leq 0.5x$ hazard: **1635 +- 125**.
- Average thinking time in those positions: **145 seconds**.
- IPR on turns of $\geq 2x$ hazard: **2345 +- 125**.
- Average thinking time in those positions: **151 seconds**.

Results are more as-expected on turns with little time budget left:

- When player has $\leq$ 180 seconds left (633 turns): **1540 +- 280**.
- Or average $\leq$ 60 seconds left to turn 40, not counting increment time: **1685 +- 200**.
- Or average 30 seconds left to turn 40, counting half the increment time: **1395 +- 425**. (In all cases, average hazard.)

Fast Chess and Player Development

- During the pandemic, I kept my model trained from 2010–2019.
- I used a player growth estimation curve devised in November 2020.
- The curve worked accurately clear thru the Budapest Olympiad.
- Notable applications: Sarayu Velpula, Hans Niemann.
- **"In-the Field" Conclusion:** Online chess and study, generally at fast paces, **was just as good for developing young minds as in-person slow chess tournaments.**
- Whether this elevated tactical ability at the expense of positional play needs further study.
- There is a shift of s and c balance with faster time controls.
- I've previously claimed evidence that online blitz is played to the same quality as in-person blitz. **Now unclear. Recent more precise calibration may allow online blitz to be 35-or-so Elo better at TT 180+1 pace. Complicated by various factors.**

Player Estimation

- Model → **Intrinsic Performance Rating** (**IPR**) for any games.
- IPR still may overdo *accuracy*, undercut *challenge created*.
- The s, c, h... tradeoff that produces a given Elo IPR value judges positional versus tactical abilities.

Questions that IPR can answer:

1. Natural growth curves for young players? & arcs for older players?
2. Are there substantial geographical variations in ratings?
3. How does skill at fast chess correlate with ratings at slow chess?
4. Has there been rating **inflation**? Is there current **deflation**?

Rating estimation bias skews linearly, but my model has ample cross-checks by which to detect and correct it. The pandemic brought a truly monstruous situation where official ratings were frozen for years...

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating (to 2637).
- Hou Yifan is 2633 but semi-inactive; next is Ju Wenjun at 2563.
- (But are current top female players more distinctly underrated?)
- Where and when does the gap begin?
- "Nature versus Nurture"—or rather **Duration of Engagement**?
- I have not found differences between these improvement factors:
  - Playing in-person chess events—versus binging online blitz.
  - Study alone—versus with a regular chess coach (online).
- What data could test a simple "10,000 hours" hypothesis?
- Perhaps: time spent on major platforms, crosstabbed by age, rating, and gender. Alas not maintained as such?
- **Q&A**, and **Thanks**.

Discussion and Q & A

[And Thanks]

[Possible extra slides for Q & A follow...optional, of course...]