

# Cheating Detection and Cognitive Modeling At Chess

## Cognitive Science Colloquium

Kenneth W. Regan<sup>1</sup>  
University at Buffalo (SUNY)

16 October, 2024

---

<sup>1</sup>With grateful acknowledgment to co-authors Guy Haworth and Tamal Biswas, students in my graduate seminars, and UB's Center for Computational Research (CCR)

# 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  $u_i$ . The main risk/reward quantity is then  $E = \sum_i p_i u_i$ . **Examples:**

- **Insurance:**  $m_i$  are risk factors; costs  $u_i$  do not influence  $p_i$ .
- **Chess:**  $m_i$  are legal moves;  $u_i$  are values given by strong chess-playing programs that objectively say how good the moves are. In my model,  $p_i$  depend on  $u_i$  per **bounded rationality**.
- **Multiple-choice tests:**  $m_i$  are possible answers to a test question,  $u_i = \text{gain/loss}$  for right/wrong answer.

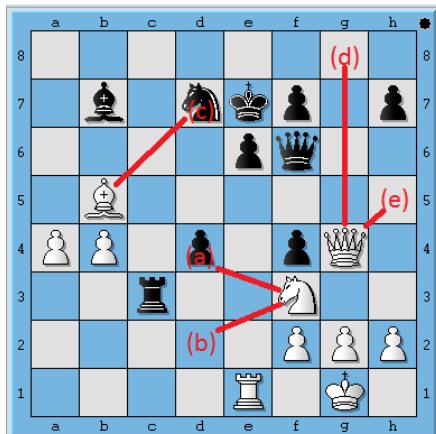
# Chess and Tests—With Partial Credits (Or LLMs?)

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

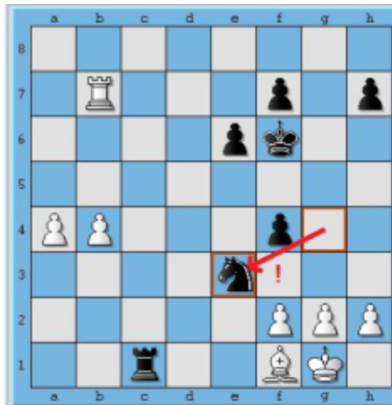
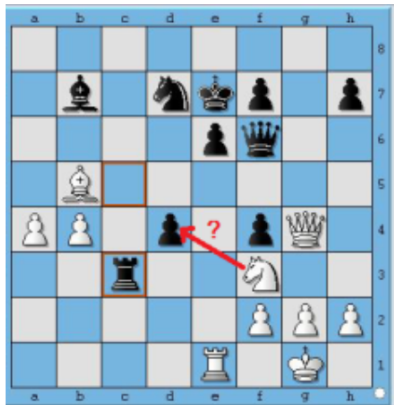
(source: itunes.apple.com)

=



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)



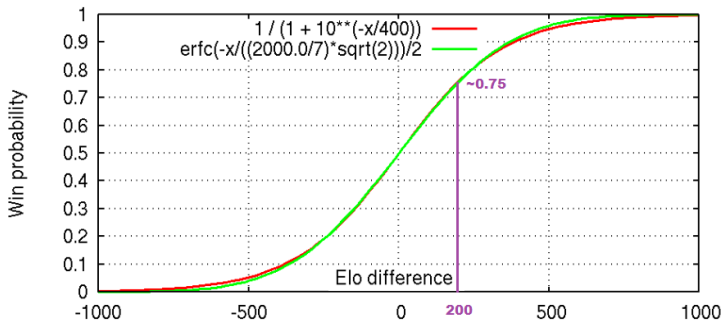
Depths...

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  $e = \frac{1}{1 + \exp(c(R_P - R_O))}$  depends only on difference to opponent's rating  $R_O$ . With  $c = (\ln 10)/400$  the curve is:



## 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 between players within 10 Elo of a marker 1025, 1050, . . . , 2775, 2800, 2825.

Wider selection below 1500 and above 2500.

- Given an Elo rating  $R$ , “central slice” gives corresponding  $s_R, c_R, h_R$ .
- Only other input is move values at various depths of search.
- Important “differentiator”: my heavily scaled version (**ASD**) of “*average centipawn loss.*”
- Other than these, **my model knows nothing about chess.**

## Log-Linear Versus Loglog-Linear Model

The generic **log-linear** model puts

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

where  $\delta_i = u_1 - u_i$ . Solved by **softmax** giving  $p_i = p_1 \exp(-\beta u_i)$ , so each  $p_i$  is represented as a **multiple** of the best-move probability  $p_1$ .

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

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

This gives  $p_i = p_1^{\exp(\beta \delta_i)}$ , so probabilities are represented as **powers** of  $p_1$ .

In place of  $\beta \delta_i$ , I have  $\left(\frac{\delta_i - h \rho_i}{s}\right)^c$ , where the “heave term”  $\rho_i$  uses the values at lower depths of search. **Why  $h$  is tightly clamped.**

*A rare bird?* Relation to *power-law* phenomena?

## How it Works

- 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.



## 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.

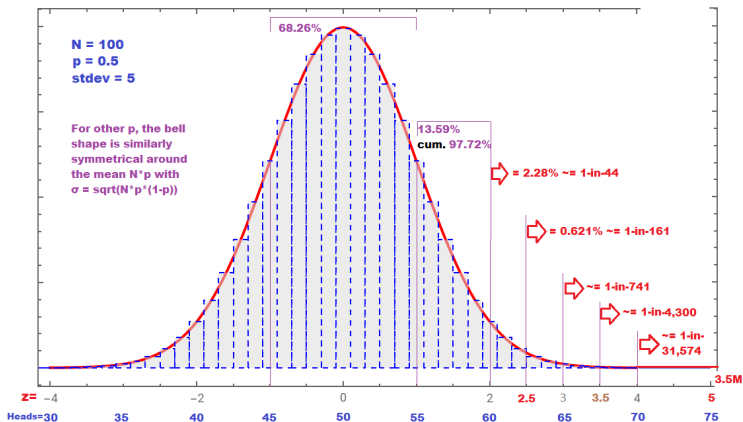
Are these grainy parameters enough to mimic human tendencies?

- 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** measures 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” ( $\sigma$ ).
- Correspond to statements of odds-against (**but see next slides**):
- “Six Sigma” ( $6\sigma$ ) means about 500,000,000–1 odds;
- $5\sigma = 3,000,000-1$ ;
- $4.75\sigma = 1,000,000-1$ ;
- $4.5\sigma = 300,000-1$ ;
- $4\sigma = 32,000-1$ ;
- $3\sigma = 740-1$ ;
- $2\sigma = 43-1$  (civil minimum standard, polling “margin of error”).

# Bell Curve and Tails



## 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.

## Evaluation Criteria and Demonstrations

- 1 Is it **safe**? That is, do its outputs conform to an expected (normal) distribution over populations that obey the null hypothesis? (Yes).
- 2 Is it **sensitive**? And are its positive results clearly pertinent to the desired inferences? (Can improve?)
- 3 How is it calibrated? Are the calibration—as well as positive results—**explainable**?
- 4 Can it be **cross-validated**? What sanity checks does it provide?
- 5 Does it model more than what its proximate application demands, so as to be robust against “mission creep”?
- 6 How can we distinguish *uncovering genuine cognitive phenomena* from *artifacts of the model*?

Show demos as time allows...

## Cognitive Concepts and Conceits

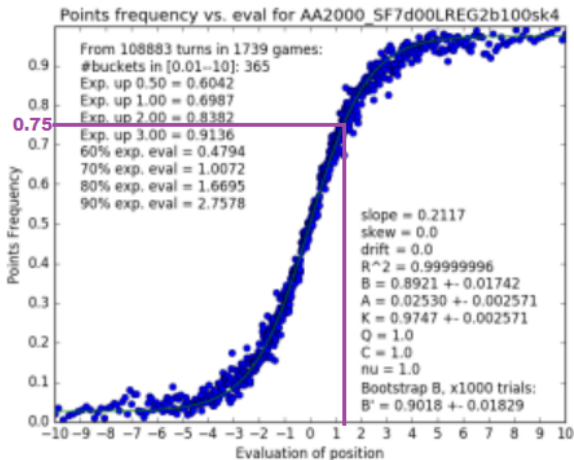
Many results in cognitive decision making come from studies that

- 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...,  
and where
- ⑤ ...reproducibility is doubtful and arduous.

The *chess angle* is to trade 1 against wealth of 2,3,4,5: lots of players and games, real competition, clear goals and metrics (Elo ratings), and not only reproducible but conducive to abundant falsifiable predictions.

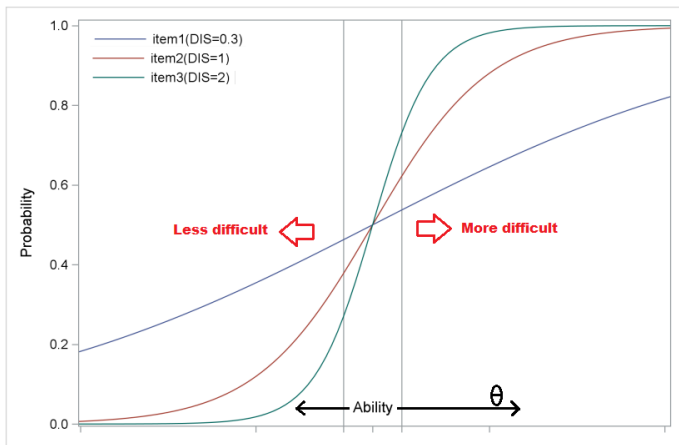
My Kahneman obit.

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

Position Value  $\longleftrightarrow$  Expectation (2000 vs. 2000)

- Similar **0.75** expectation when up 1.30 vs. equal-rated player.
- Complication: **dependence** on rating itself.

# Item-Response Theory (IRT [source](#))



- Horizontal axis governs **difficulty** in relation to  $\theta = \text{ability}$ .
- Slope at  $y = 0.5$  *correctness rate* is the **discrimination** factor.



## Defining Difficulty

- For any *fixed* aptitude level  $\theta$ , *difficulty*  $\approx$  *expected points loss*.
- In chess, this is our  $E_L = \sum_i p_i(u_1 - u_i) = \sum_i p_i \delta_i$ .
- Call this expected loss the **hazard**.
- Depends on rating because the probabilities  $p_i$  projected by my model depend on rating  $R$ .
- My model divides out dependence on  $R$ . “Expectation Weights, Normalized” (EWN).
- *Technote*: In a **log-linear** model, with  $-\log p_i \sim u_i$ , we get

$$E_L \sim \sum_i p_i \log(1/p_1) - \sum_i p_i \log(1/p_i) = \log\left(\frac{1}{p_1}\right) - H,$$

where  $H$  is **entropy**. But my model is not log-linear.

- How well does hazard—normalized over aptitude—work as a measure of difficulty?

## A Philosophical Issue

Should a grading metric  $\mu$  expect to assess lower performance on more-difficult questions, or should it show a *constancy of signal  $\theta$*  across all types of questions?

- I typically categorize questions as A-level, B-level, C-level, D-level.
- Ideal distribution: 20%,30%,30%,20% averaging **2.5** difficulty.
- Overall threshold for A: grading score  $\mu \geq 90\%$ .
- Getting 60% on the A-level questions puts you on-track, even though 60% by itself is C-range (or worse).
- Thus simple  $\mu$  does not give constant signal—it needs context.
- Should we define “A-level” etc. in each category? ( $\approx$  *curving*).

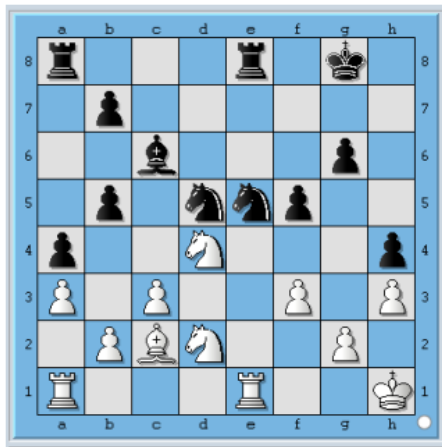
*Raw metrics* like T1, EV, ASD should not give constancy of signal.

**How about IPR?**

## IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average **2080**.
- Focus on **2000–2200**. Analysis by Stockfish 11 in **EWN** mode.
- IPR overall: **2125** +/- **40**. Broken down according to [dis-]advantage:
  - 1–2 pawns behind: **2170** +/- **105**; worse: **2065** +/- **110**.
  - 1–2 pawns ahead: **2085** +/- **120**; better: **2020** +/- **155**
  - Within 1.00 of equal: **2145** +/- **45**; within 0.50: **2125** +/- **65**.
- Reasonable constancy of signal.
- But on positions with  $\geq 1.5$  times normal hazard: **2255** +/- **65**.
- With  $\geq 2x$  hazard: **2170** +/- **115**. Could be consistent. **But—**
- Positions of of  $0.5x$  or lower hazard: **1800** +/- **180**.
- Not constancy of signal.
- Low-hazard positions either have an obvious best move or many good moves.

# Example: 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  $E_L$ , high entropy  $H$ . (Niemann lost.)

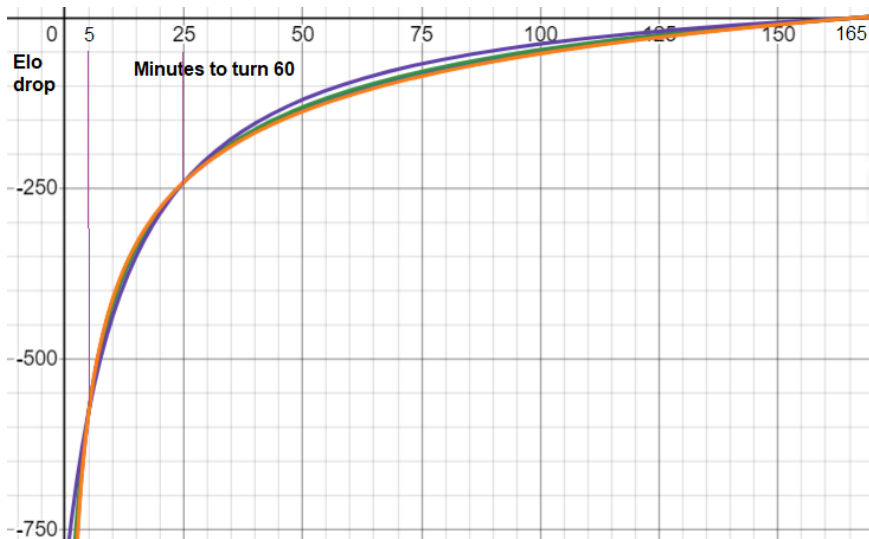
## Aspects of Difficulty (Besides Hazard)

- 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.** *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.**
  - *Can't project ahead of time* (owing to non-book  $\equiv$  non-repeatable).
  - But certainly directly captures the human *experience* of difficulty.
- 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.
- **Blitz** means under 10 minutes, most common is 3 minutes + 2-second increment, which gives **5** minutes—and so approximates old-school 5-minute chess on analog clocks.
- For 25-minute Rapid, I measure **240** reduction in quality per IPR.
- For 5-minute Blitz, **575** lower. (Error bars for both are about  $\pm 25$ .)

# Time-Quality Curves (whole graph)



## Predicated on Time Spent For a Move

Staying with players rated 2000 to 2200 at the World Senior Team Ch.

- 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).



## 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.)

## Enter Entropy

Students in my CSE702 graduate seminar proposed a measure  $H_U$  of entropy that uses only the move utilities  $u_i$ , not the projected probabilities  $p_i$  (nor their logs). Avoids the rating feedback loop.

- Average  $H_U = 2.57$ .
- Turns with  $H_U \leq 2$ : avg. time used **88 sec.**, IPR **2405 +- 100**.
- Turns with  $H_U \leq 1.5$ : avg. time used **72 sec.**, IPR **2485 +- 130**.
- Turns with  $H_U \leq 1$ : avg. time used **56 sec.**, IPR **2645 +- 165**  
(lower hazard too).
- Turns with  $H_U \leq 0.5$ : avg. time used **40 sec.**, IPR **2580 +- 255**  
(much lower hazard).
- Turns with  $H_U \geq 3$ : time used **252 sec.**, IPR **2000 +- 35**.
- Turns with  $H_U \geq 3.5$  (702 pos.): time **312 sec.**, IPR **1965 +- 110**.
- (No position has  $H_U \geq 3.8$ . All cases have close to mean hazard.)
- High entropy correlates well with (human experience of) difficulty.
- Much more work to do...

## Discussion and Q & A

[And Thanks]

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

## Some Accompanying Stances

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim.
- Concern: Data modelers in less-extreme settings **satisfice**.
- That is, their models are designed up to one particular goal but don't explore much of the harder adjacent metaspace.
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of [this study](#).
- **Cross-Validation**...one point of which is:
- How can we distinguish *uncovering genuine cognitive phenomena* from *artifacts of the model*?

## Some Cognitive Nuggets

- ① Dimensions of Strategy and Tactics (and Depth of Thinking).
  - But wait—the model has no information specific to chess...
  - Brain seems to register changes in move values as depth increases.
- ② Machine-Like Versus Human Play
  - Garry Kasparov, as a 2012 Alan Turing Centennial test, distinguished 5 games played by human 2200-level masters from 5 games by engines “stopped down” to 2200 level.
- ③ Relationship to Multiple-Choice Tests (with partial credits)
  - “Solitaire Chess” feature often gives part credits.
  - Large field of **Item Response Theory** (IRT).

## 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:

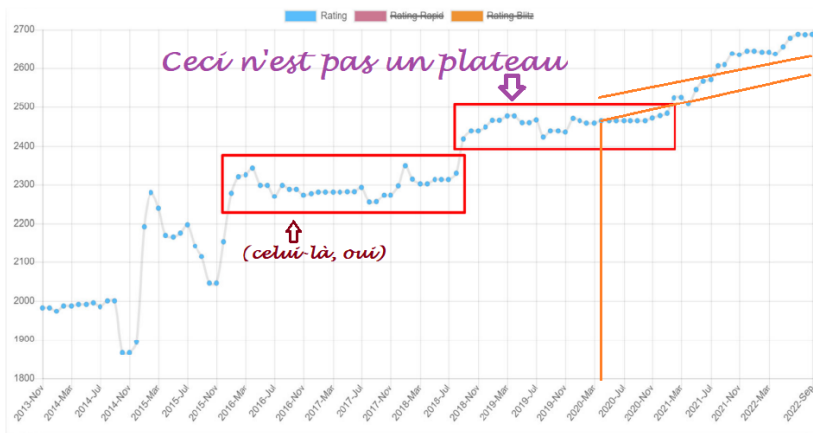
- ① Natural growth curves for young players? & arcs for older players?
- ② Are there substantial **geographical variations** in ratings?
- ③ How does skill at fast chess correlate with ratings at slow chess?
- ④ Has there been rating **inflation**? Is there current **deflation**?

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

## Rating Lag—Natural Versus Pandemic-Caused

- **The #1 scientific role I've played since the pandemic has been estimating the true skill growth of young players.**
- Has perforce been **post-normal science**.
- My “back of the envelope” formula held up over two years with only one small revision for preteens.
- Revision in Oct. 2022 to curtail projections past Elo 2000 level.
- Would have been more “normal” if comprehensive studies of the career arcs (measured by Elo rating) of young players were to hand.
- Lack of such studies exposed by the controversy over Hans Niemann's rise from 2465 Elo to 2700.
- Show **this GLL article** including example of Ms. Sarayu Velpula.
- **Velpula in current Indian Women's Championship...**

# Hans Niemann: Platform or Plateau?





## 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**.