Time and Difficulty
Artificial Intelligence and Sustainable Computing (AISC 2024)

Kenneth W. Regan¹
University at Buffalo (SUNY)

13 July, 2024

¹With 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 $m_1, m_2, \ldots, m_j, \ldots$
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes \( m_1, m_2, \ldots, m_j, \ldots \)
- Assigns to each \( m_j \) a probability \( p_j \).
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each $m_j$ a probability $p_j$.
- Projects risk/reward quantities associated to the outcomes.

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, $u_i = \text{gain/loss for right/wrong answer}$.
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- 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.
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes \( m_1, m_2, \ldots, m_j, \ldots \)
- 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:
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes \( m_1, m_2, \ldots, m_j, \ldots \).
- 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 \( p_i \).
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- 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.
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- 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.
A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions, each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- 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, $u_i = \text{gain/loss for right/wrong answer.}$
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)
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.
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.
A Difficult Trap (Kramnik-Anand, 2008 WC)
Aptitude—Via Elo Grades (calculator)

- Named for Arpad Elo, number $R_P$ rates skill of player $P$. 

$$
e = \frac{1}{1+\exp\left(c(R_P - R_O)\right)}$$
Expectation depends only on difference to opponent's rating $R_O$. With $c = \frac{\ln 10}{400}$ the curve is:
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:
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.
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:
Position Value $\leftrightarrow$ Expectation (2000 vs. 2000)

From 108883 turns in 1739 games:
#buckets in [0.01–10]: 365
Exp. up 0.50 = 0.6042
Exp. up 1.00 = 0.6987
Exp. up 2.00 = 0.8382
Exp. up 3.00 = 0.9136
60% exp. eval = 0.4794
70% exp. eval = 1.0072
80% exp. eval = 1.6695
90% exp. eval = 2.7578

slope = 0.2117
skew = 0.0
drift = 0.0
$R^2 = 0.99999996$
B = 0.8921 +/- 0.01742
A = 0.02530 +/- 0.002571
K = 0.9747 +/- 0.002571
Q = 1.0
C = 1.0
nu = 1.0
Bootstrap B, x1000 trials:
B* = 0.9018 +/- 0.01829
• Similar *0.75* expectation when up 1.30 vs. equal-rated player.
• 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.
Horizontal axis governs **difficulty** in relation to $\theta = \text{ability}$.
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 $E_L = P_i p_i (u_1 - u_i) = P_i p_i \delta_i$. 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).

Technotes: In a log-linear model, $-\log p_i \sim u_i$. Then $E_L \sim P_i p_i \log(1/p_1) - P_i p_i \log(1/p_i) = \log(1/p_1) - H$ where $H$ is entropy. However, my model is double-log linear: $\log p_i \log p_1 \sim \exp(\delta_i)$. Why double-log works and single-log fails.

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

- For any fixed aptitude level $\theta$, difficulty $\approx$ expected points loss.
Defining Difficulty

- For any fixed aptitude level \( \theta \), difficulty \( \approx \) expected points loss.
- In chess, this is our \( E_L = \sum_i p_i(u_1 - u_i) = \sum_i p_i \delta_i \).
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.
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 $R$. 

\[ E_L = \sum_i p_i(u_1 - u_i) = \sum_i p_i \delta_i. \]
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 $R$.
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 $R$.
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 $R$.
- Technotes: In a log-linear model, $-\log p_i \sim u_i$. 
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 $R$.
- Technotes: In a log-linear model, $-\log p_i \sim u_i$.
- Then $E_L \sim \sum_i p_i \log(1/p_1) - \sum_i p_i \log(1/p_i) = \log(\frac{1}{p_1}) - H$ where $H$ is entropy.
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).

Technotes: In a log-linear model, $-\log p_i \sim u_i$.

Then $E_L \sim \sum_i p_i \log(1/p_1) - \sum_i p_i \log(1/p_i) = \log(\frac{1}{p_1}) - H$ where $H$ is entropy.

However, my model is double-log linear: $\frac{\log p_i}{\log p_1} \sim \exp(\delta_i)$. 
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 $R$.
- Technotes: In a **log-linear** model, $-\log p_i \sim u_i$.
- Then $E_L \sim \sum_i p_i \log(1/p_1) - \sum_i p_i \log(1/p_i) = \log(\frac{1}{p_1}) - H$ where $H$ is entropy.
- **However**, my model is **double-log linear**: $\frac{\log p_i}{\log p_1} \sim \exp(\delta_i)$.
- Why double-log works and single-log fails.
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 $R$.
- *Technote*: In a **log-linear** model, $-\log p_i \sim u_i$.
- Then $E_L \sim \sum_i p_i \log(1/p_1) - \sum_i p_i \log(1/p_i) = \log(\frac{1}{p_1}) - H$ where $H$ is entropy.
- *However*, my model is **double-log linear**: $\frac{\log p_i}{\log p_1} \sim \exp(\delta_i)$.
- Why double-log works and single-log fails.
- How well does hazard—normalized over aptitude—work as a measure of difficulty?
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?
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
- Overall threshold for A: 90%.
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
- Overall threshold for A: 90%.
- Getting 60% on the A-level questions puts you on-track, even though 60% by itself is C-range (or worse).
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
- Overall threshold for A: 90%.
- Getting 60% on the A-level questions puts you on-track, even though 60% by itself is C-range (or worse).
- Thus the simple grading score $\mu$ does not give constant signal—it needs context.
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
- Overall threshold for A: 90%.
- Getting 60% on the A-level questions puts you on-track, even though 60% by itself is C-range (or worse).
- Thus the simple grading score $\mu$ does not give constant signal—it needs context.
- Should we use metrics that say “A-level” etc. in each category?
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 design exams to have 20% A-level questions, 30% B-level, 30% C-level, 20% D-level.
- Overall threshold for A: 90%.
- Getting 60% on the A-level questions puts you on-track, even though 60% by itself is C-range (or worse).
- Thus the simple grading score $\mu$ does not give constant signal—it needs context.
- Should we use metrics that say “A-level” etc. in each category? (Like curving).
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but scaled down when one side has advantage.
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but *scaled* down when one side has advantage.
  - Called **ACPL** for *average centipawn loss* without scaling.

My **Intrinsic Performance Rating** (**IPR**) metric fits parameters $s$ for “sensitivity” ($\sim$ strategic ability), and $c$ for “consistency” (in surviving tactical minefields) to give the closest **Virtual Player** $P(s, c)$ on any set of games. Then trained correspondence $(s, c) \rightarrow R$ gives IPR as an Elo rating. Should give constancy of signal...but...
The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but *scaled* down when one side has advantage.
  - Called **ACPL** for *average centipawn loss* without scaling.

All should vary with difficulty, hence not give constancy of signal.
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but scaled down when one side has advantage.
  - Called **ACPL** for average centipawn loss without scaling.

All should vary with difficulty, hence not give constancy of signal.

- My **Intrinsic Performance Rating (IPR)** metric fits parameters
  - *s* for “sensitivity” (∼ strategic ability), and
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but *scaled* down when one side has advantage.
  - Called **ACPL** for *average centipawn loss* without scaling.

All should vary with difficulty, hence not give constancy of signal.

- **My Intrinsic Performance Rating (IPR)** metric fits parameters
  - $s$ for “sensitivity” ($\sim$ strategic ability), and
  - $c$ for “consistency” (in surviving tactical minefields)

  to give the closest *Virtual Player* $P(s, c)$ on any set of games.
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but *scaled* down when one side has advantage.
  - Called **ACPL** for *average centipawn loss* without scaling.

All should vary with difficulty, hence not give constancy of signal.

- My **Intrinsic Performance Rating (IPR)** metric fits parameters
  - *s* for “sensitivity” (∼ strategic ability), and
  - *c* for “consistency” (in surviving tactical minefields)
  to give the closest *Virtual Player* $P(s, c)$ on any set of games.
- Then trained correspondence $(s, c) \rightarrow R$ gives IPR as an Elo rating.
Model and Metrics

The following “raw metrics” on series of games are used generally:

- **T1-match**: Agreement with the move listed first by the computer.
- **EV-match**: Includes moves of equal-optimal value not listed first.
- **ASD**: Average difference in value from inferior moves (over all positions), but *scaled* down when one side has advantage.
  - Called **ACPL** for *average centipawn loss* without scaling.

All should vary with difficulty, hence not give constancy of signal.

- **My Intrinsic Performance Rating (IPR)** metric fits parameters
  - $s$ for “sensitivity” ($\sim$ strategic ability), and
  - $c$ for “consistency” (in surviving tactical minefields)

  to give the closest **Virtual Player** $P(s, c)$ on any set of games.

- Then trained correspondence $(s, c) \rightarrow R$ gives IPR as an Elo rating.
- Should give constancy of signal...but...
How Accurate Are Model Projections?

Internal evidence that it gives $\approx (1 + \epsilon)$ relative error with $\epsilon \approx 0.04$ for most rating levels.
How Accurate Are Model Projections?

Internal evidence that it gives \( \approx (1 + \epsilon) \) relative error with \( \epsilon \approx 0.04 \) for most rating levels. Means it supports betting on chess moves with only 5% “vig” to avoid arbitrage.
How Accurate Are Model Projections?

Internal evidence that it gives $\approx (1 + \epsilon)$ relative error with $\epsilon \approx 0.04$ for most rating levels. Means it supports betting on chess moves with only 5% “vig” to avoid arbitrage. (Except for bets against clear-best moves.)
IPR and Hazard (World Senior Teams 2024)

Older players, established ratings (but deflated), average 2080.


IPR overall: 2125 ± 40.

Broken down according to disadvantage:
- 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 ≥1.5 times normal hazard: 2255 ± 65.

With ≥2x hazard: 2170 ± 115.

Could be consistent.

But—

Positions of 0.5 or lower hazard: 1800 ± 180.

Not constancy of signal.

Low-hazard positions either have an obvious best move or many good moves.
Older players, established ratings (but deflated), average 2080.

**IPR and Hazard (World Senior Teams 2024)**

- IPR overall: 2125 ± 40.
- Broken down according to 
  - 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 ≥ 1.5 times normal hazard: 2255 ± 65.
- With ≥ 2x hazard: 2170 ± 115.
- Could be consistent.
- But—positions 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.
Older players, established ratings (but deflated), average 2080.
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**.
Older players, established ratings (but deflated), average 2080.


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.
IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average 2080.
- 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.
IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average 2080.


- 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 ≥ 1.5 times normal hazard:
Older players, established ratings (but deflated), average 2080.


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 ≥ 1.5 times normal hazard: 2255 +- 65.
Older players, established ratings (but deflated), average 2080.


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.
IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average 2080.
- 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 ≥ 1.5 times normal hazard: 2255 +- 65.
- With ≥ 2x hazard: 2170 +- 115. Could be consistent. But—
IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average 2080.
- 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 ≥ 1.5 times normal hazard: 2255 +- 65.
- With ≥ 2x hazard: 2170 +- 115. Could be consistent. But—
- Positions of of 0.5x or lower hazard: 1800 +- 180.
IPR and Hazard (World Senior Teams 2024)

- Older players, established ratings (but deflated), average 2080.
- 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 ≥ 1.5 times normal hazard: 2255 +- 65.
- With ≥ 2x hazard: 2170 +- 115. Could be consistent. But—
- Positions of of 0.5x or lower hazard: 1800 +- 180.
- Not constancy of signal.
Older players, established ratings (but deflated), average 2080.


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.
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 the best move or avoid a trap. Expressly modeled—e.g., to project the trap for Kramnik.


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 ≡ 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.
1. Needing deep cogitation to find best move or avoid a trap. 
   *Expressly modeled—e.g. to project the trap for Kramnik.*
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.**
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.*
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.*
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.

4. 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.

5. Carow and Witzig [CW, Feb. 2024] consider all the above, but strive for human-chess based measures.
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.*
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.*
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.
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 ≡ non-repeatable).
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 ≡ non-repeatable).
   - But certainly directly captures the human *experience* of difficulty.
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 ≡ non-repeatable).
   - But certainly directly captures the human experience of difficulty.
5. Question is inherently complex or taxing.
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 ≡ non-repeatable).
   - But certainly directly captures the human *experience* of difficulty.

5. Question is inherently complex or taxing.
   - How to measure this internally?
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 chwess engine to complete analysis to depth (say) 24.
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 ≡ 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 chwess 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 and Difficulty

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 ±25.)
Time and Difficulty

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.
Time and Difficulty

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.
Time and Difficulty

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.
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).
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.
Time and Difficulty

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.
**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.
Time and Difficulty

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.
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 ±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.

"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).
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:

- Positions with 5–10 minutes consumed:

- Using 10–15 minutes (705 positions):

- Using ≥ 15 minutes (371 positions):

"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).
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: $2860 \pm 75$.

At most 10 seconds: $3235 \pm 90$.

Starting at turn 16 rather than 9: $3220 \pm 100$.

At most 5 seconds (sample size 605): $3230 \pm 160$.

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

Positions with 5–10 minutes consumed: $1460 \pm 85$.

Using 10–15 minutes (705 positions): $1235 \pm 170$.

Using $\geq 15$ minutes (371 positions): $1410 \pm 205$.

"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).
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**.

"Thinking Is Bad For You." (At least it’s a bad sign...)
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**.

"Thinking Is Bad For You."
(At least it's a bad sign...)
Time and Difficulty

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—?
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:
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.
Time and Difficulty

**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: \(2860 \pm 75\).
- At most **10 seconds**: \(3235 \pm 90\).
- Starting at turn 16 rather than 9: \(3220 \pm 100\).
- At most **5 seconds** (sample size 605): \(3230 \pm 160\).

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

- Positions with 5–10 minutes consumed: \(1460 \pm 85\).
- Using 10–15 minutes (705 positions): \(1235 \pm 170\).
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 ≥ 15 minutes (371 positions): 1410 ± 205.

"Thinking Is Bad For You." (At least it’s a bad sign...)
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: \(2860 \pm 75\).
- At most 10 seconds: \(3235 \pm 90\).
- Starting at turn 16 rather than 9: \(3220 \pm 100\).
- At most 5 seconds (sample size 605): \(3230 \pm 160\).

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

- Positions with 5–10 minutes consumed: \(1460 \pm 85\).
- Using 10–15 minutes (705 positions): \(1235 \pm 170\).
- Using \(\geq 15\) minutes (371 positions): \(1410 \pm 205\).
- “Thinking Is Bad For You.”
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: $2860 \pm 75$.
- At most $10$ seconds: $3235 \pm 90$.
- Starting at turn 16 rather than 9: $3220 \pm 100$.
- At most $5$ seconds (sample size 605): $3230 \pm 160$.

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

- Positions with 5–10 minutes consumed: $1460 \pm 85$.
- Using 10–15 minutes (705 positions): $1235 \pm 170$.
- Using $\geq 15$ minutes (371 positions): $1410 \pm 205$.
- “Thinking Is Bad For You.” (At least it’s a bad sign...)
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).
Switching to Komodo 13.3 in place of Stockfish 11 as analyzing engine:

- Overall IPR of Elo 2000-to-2200 players: $2175 \pm 35$.
- Average thinking time over all moves (turns 9–60): 181 seconds.
- IPR on turns of $\leq 0.5$ hazard: $1635 \pm 125$.
- Average thinking time in those positions: 145 seconds.
- IPR on turns of $\geq 2$x hazard: $2345 \pm 125$.
- 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): $1540 \pm 280$.
- Or average $\leq 60$ seconds left to turn 40, not counting increment time: $1685 \pm 200$.
- Or average 30 seconds left to turn 40, counting half the increment time: $1395 \pm 425$.

(In all cases, average hazard.)
Hazard Vs. Time—and Time Left

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

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: \textbf{2175} \textpm 35.
- Average thinking time over all moves (turns 9–60): \textbf{181 seconds}.
Switching to Komodo 13.3 in place of Stockfish 11 as analyzing engine:

- Average thinking time over all moves (turns 9–60): 181 seconds.
- IPR on turns of ≤ 0.5x hazard: 1635 +- 125.
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 ≤ 0.5x hazard: **1635 ± 125**.
- Average thinking time in those positions: **145 seconds**.
Switching to Komodo 13.3 in place of Stockfish 11 as analyzing engine:

- Average thinking time over all moves (turns 9–60): 181 seconds.
- IPR on turns of ≤ 0.5x hazard: 1635 +- 125.
- Average thinking time in those positions: 145 seconds.
- IPR on turns of ≥ 2x hazard: 2345 +- 125.
Switching to Komodo 13.3 in place of Stockfish 11 as analyzing engine:

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

Results are more as-expected on turns with little time budget 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**.
Hazard Vs. Time—and Time Left

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

- Average thinking time over all moves (turns 9–60): 181 seconds.
- IPR on turns of ≤ 0.5x hazard: 1635 +- 125.
- Average thinking time in those positions: 145 seconds.
- IPR on turns of ≥ 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 ≤ 180 seconds left (633 turns): 1540 +- 280.
- Or average ≤ 60 seconds left to turn 40, not counting increment time: 1685 +- 200.
Hazard Vs. Time—and Time Left

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

- Overall IPR of Elo 2000-to-2200 players: \(2175 \pm 35\).
- Average thinking time over all moves (turns 9–60): 181 seconds.
- IPR on turns of \(\leq 0.5x\) hazard: \(1635 \pm 125\).
- Average thinking time in those positions: 145 seconds.
- IPR on turns of \(\geq 2x\) hazard: \(2345 \pm 125\).
- 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): \(1540 \pm 280\).
- Or average \(\leq 60\) seconds left to turn 40, not counting increment time: \(1685 \pm 200\).
- Or average 30 seconds left to turn 40, counting half the increment time: \(1395 \pm 425\). (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).
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.
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 $H_U = 2.57$. 
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...
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.
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).

(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...
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 $2405 \pm 100$.
- Turns with $H_U \leq 1.5$: avg. time used 72 sec., IPR $2485 \pm 130$.
- Turns with $H_U \leq 1$: avg. time used 56 sec., IPR $2645 \pm 165$ (lower hazard too).
- Turns with $H_U \leq 0.5$: avg. time used 40 sec., IPR $2580 \pm 255$ (much lower 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.
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.
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.)
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...
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...]
Many results in cognitive decision making come from studies that are well-targeted to the concept and hypothesis, but...
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...
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...
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
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.
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 Accompanying Stances

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim.
Some Accompanying Stances

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim.
- Concern: Data modelers in less-extreme settings satisfy.
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.
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.
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...
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 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
Some Cognitive Nuggets

- Dimensions of Strategy and Tactics
Some Cognitive Nuggets

1. Dimensions of Strategy and Tactics (and Depth of Thinking).
Some Cognitive Nuggets

1. Dimensions of Strategy and Tactics (and Depth of Thinking).
   - But wait—the model has no information specific to chess...
Some Cognitive Nuggets

- 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.
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
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.
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)
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.
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).
Time and Difficulty

Player Development

Note low Montreal 1979 IPRs. Even further deflation at the 1986 Men's and Women's Olympiads in Dubai.

"Today's players deserve their ratings." Is human performance at chess improving as with physical sports?...because of computers?

Growth Curves of Improving (Young) Players. How To Manage Time Budget (basically, follow V. Anand!).
Rating Inflation? Deflation?
Rating Inflation? Deflation?

- Note low Montreal 1979 IPRs.
Rating Inflation? Deflation?

- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
Rating Inflation? Deflation?

- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
- “Today’s players deserve their ratings.”
Rating Inflation? Deflation?

- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
- “Today’s players deserve their ratings.”
- Is human performance at chess improving as with physical sports?
Rating Inflation? Deflation?

- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
- “Today’s players deserve their ratings.”
- Is human performance at chess improving as with physical sports?...because of computers?
Player Development

Rating Inflation? Deflation?
- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
- “Today’s players deserve their ratings.”
- Is human performance at chess improving as with physical sports? ...because of computers?

Growth Curves of Improving (Young) Players.
Rating Inflation? Deflation?
- Note low Montreal 1979 IPRs.
- Even further deflation at the 1986 Men’s and Women’s Olympiads in Dubai.
- “Today’s players deserve their ratings.”
- Is human performance at chess improving as with physical sports? ...because of computers?

Growth Curves of Improving (Young) Players.

How To Manage Time Budget (basically, follow V. Anand!).
Cancer and Covid (= in-person and online chess)

Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people... and get a positive. What are the odds that you have the cancer? Not the same as the odds that any one test result is wrong.

Consider giving the test to 5,000 people, including yourself. Among them, 1 has the cancer; expect that result to be positive. But we can also expect about 100 false positives. All you know at this point is: you are one of 101 positives. So the odds are still 100-1 against your having the cancer. The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way.

Need a "Second Opinion." IMPHO, 1-in-5,000 ≈ frequency of cheating in-person. A positive from a "98%" test is like getting \( z = 2.05 \). Not enough.

In a 500-player Open, you should see ten such scores.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98\% accurate for a cancer that affects 1-in-5,000 people...

- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a "Second Opinion."

- IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.

- A positive from a "98\%" test is like getting \( z = 2.05 \).

- Not enough.

- In a 500-player Open, you should see ten such scores.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. What are the odds that you have the cancer?

IMPHO, 1-in-5,000 \approx frequency of cheating in-person.

A positive from a “98%” test is like getting $z = 2.05$.

Not enough.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.

IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.
A positive from a "98%" test is like getting $z = 2.05$. Not enough.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, 1 has the cancer; expect that result to be positive.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects **1-in-5,000** people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
Cancer and Covid (= in-person and online chess)

Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...

...and get a positive. What are the odds that you have the cancer?

Not the same as the odds that any one test result is wrong.

Consider giving the test to 5,000 people, including yourself.

- Among them, 1 has the cancer; expect that result to be positive.
- But we can also expect about 100 false positives.
- All you know at this point is: you are one of 101 positives.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*

Not the same as the odds that any one test result is wrong.

Consider giving the test to 5,000 people, including yourself.

- Among them, **1** has the cancer; expect that result to be positive.
- But we can also expect about **100** false positives.
- All you know at this point is: you are **one** of **101** positives.

- So the odds are still **100-1 against** your having the cancer.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, 1 has the cancer; expect that result to be positive.
  - But we can also expect about 100 false positives.
  - All you know at this point is: you are *one* of 101 positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. What are the odds that you have the cancer?
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, 1 has the cancer; expect that result to be positive.
  - But we can also expect about 100 false positives.
  - All you know at this point is: you are one of 101 positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a “Second Opinion.”
Cancer and Covid (in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. What are the odds that you have the cancer? Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, 1 has the cancer; expect that result to be positive.
  - But we can also expect about 100 false positives.
  - All you know at this point is: you are one of 101 positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a “Second Opinion.”
- IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is 98% accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, 1 has the cancer; expect that result to be positive.
  - But we can also expect about 100 false positives.
  - All you know at this point is: you are one of 101 positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a “Second Opinion.”
- IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.
- A positive from a “98%” test is like getting $z = 2.05$. Not enough.
Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one of 101** positives.
- So the odds are still **100-1 against** your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a “Second Opinion.”
- IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.
- A positive from a “98%” test is like getting $z = 2.05$. *Not enough.*
- In a 500-player Open, you should see **ten such scores**.
The 99.993% Test

Suppose our cancer test were 600 times more accurate: 1-in-30,000 error. That's the face-value error rate claimed by a $z = 4$ result. Still 1-in-6 chance of false positive among 5,000 people. (This is really how a "second opinion" operates in practice.) If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural. Still not comfortable satisfaction of the result being unnatural. IMHO, the interpretation of CAS comfortable-satisfaction range of final odds determination is 99%–99.9% confidence. Target confidence should depend on gravity of consequences. (CAS) Sweet spot IMHO is 99.5%, meaning 1-in-200 ultimate chance of wrong decision. Same criterion used by Decision Desk HQ to "call" US elections. Higher stringency cuts against timely public service.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: 1-in-30,000 error.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a $z = 4$ result.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: \textbf{1-in-30,000} error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still \textbf{1-in-6} chance of false positive among 5,000 people.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: 1-in-30,000 error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still **1-in-6** chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then **1-in-60** chance of the result being natural.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: \textbf{1-in-30,000} error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still \textbf{1-in-6} chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then \textbf{1-in-60} chance of the result being natural.
- Still not \textbf{comfortable satisfaction} of the result being unnatural.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a \( z = 4 \) result.
- Still **1-in-6** chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then **1-in-60** chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still **1-in-6** chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then **1-in-60** chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still **1-in-6** chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then **1-in-60** chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
- Sweet spot IMHO is **99.5%**, meaning **1-in-200** ultimate chance of wrong decision.
Suppose our cancer test were 600 times more accurate: 1-in-30,000 error.

That’s the face-value error rate claimed by a $z = 4$ result.

Still 1-in-6 chance of false positive among 5,000 people.

(This is really how a “second opinion” operates in practice.)

If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.

Still not **comfortable satisfaction** of the result being unnatural.

IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is 99%–99.9% confidence.

Target confidence should depend on gravity of consequences. (CAS)

Sweet spot IMHO is 99.5%, meaning 1-in-200 ultimate chance of wrong decision. Same criterion used by **Decision Desk HQ** to “call” US elections.
The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That’s the face-value error rate claimed by a $z = 4$ result.
- Still **1-in-6** chance of false positive among 5,000 people.
- (This is really how a “second opinion” operates in practice.)
- If the entire world were a 500-player Open, then **1-in-60** chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
- Sweet spot IMHO is **99.5%**, meaning **1-in-200** ultimate chance of wrong decision. Same criterion used by Decision Desk HQ to “call” US elections.
- Higher stringency cuts against timely public service.
Covid in Non-Surge and Surge Times

Now suppose the factual positivity rate is 1-in-50. We still have about 100 false positives, but now also 100 factual positives. A positive from a 98% test is here a 50-50 coinflip. But a negative is good: Only 2 false negatives will expect to come from the 100 dangerous people. From the 4,900 safe people, about 4,800 true negatives. Odds that your negative is false are 2,400-to-1 against. Fine to be on a plane. What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50. Now suppose the factual positivity rate is 20%. Can we do this in our heads?
Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is 1-in-50.

A positive from a 98% test is here a 50-50 coinflip. But a negative is good: Only 2 false negatives will expect to come from the 100 dangerous people. From the 4,900 safe people, about 4,800 true negatives. Odds that your negative is false are 2,400-to-1 against. Fine to be on a plane.

What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50.

Now suppose the factual positivity rate is 20%. Can we do this in our heads?
Now suppose the factual positivity rate is 1-in-50.

We still have about 100 false positives, but now also 100 factual positives.
Now suppose the factual positivity rate is 1-in-50.

We still have about 100 false positives, but now also 100 factual positives.

A positive from a 98% test is here a 50-50 coinflip.
Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is 1-in-50.
- We still have about 100 false positives, but now also 100 factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
Now suppose the factual positivity rate is **1-in-50**.

We still have about **100** false positives, but now also **100** factual positives.

A positive from a 98% test is here a 50-50 coinflip.

But a negative is *good*:

- Only 2 false negatives will expect to come from the **100** dangerous people.
Now suppose the factual positivity rate is **1-in-50**.

We still have about **100** false positives, but now also **100** factual positives.

A positive from a 98% test is here a 50-50 coinflip.

But a negative is *good*:

- Only 2 false negatives will expect to come from the **100** dangerous people.
- From the **4,900** safe people, about **4,800** true negatives.
Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is 1-in-50.
- We still have about 100 false positives, but now also 100 factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is good:
  - Only 2 false negatives will expect to come from the 100 dangerous people.
  - From the 4,900 safe people, about 4,800 true negatives.
  - Odds that your negative is false are 2,400-to-1 against.
Now suppose the factual positivity rate is **1-in-50**.

We still have about **100** false positives, but now also **100** factual positives.

A positive from a 98% test is here a 50-50 coinflip.

But a negative is *good*:

- Only 2 false negatives will expect to come from the **100** dangerous people.
- From the **4,900** safe people, about **4,800** true negatives.
- Odds that your negative is false are **2,400-to-1** against.

*Fine to be on a plane.*
Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is 1-in-50.
- We still have about 100 false positives, but now also 100 factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
  - Only 2 false negatives will expect to come from the 100 dangerous people.
  - From the 4,900 safe people, about 4,800 true negatives.
  - Odds that your negative is false are 2,400-to-1 against.
- *Fine to be on a plane.* What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50.
Now suppose the factual positivity rate is 1-in-50.

We still have about 100 false positives, but now also 100 factual positives.

A positive from a 98% test is here a 50-50 coinflip.

But a negative is *good*:
- Only 2 false negatives will expect to come from the 100 dangerous people.
- From the 4,900 safe people, about 4,800 true negatives.
- Odds that your negative is false are 2,400-to-1 against.

*Fine to be on a plane.* What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50.

Now suppose the factual positivity rate is 20%. Can we do this in our heads?
Suppose we get $z = 4$ in online chess with adult cheating rate 2%. Out of 30,000 people:

1 false positive result.
600 factual positives.
So $600 - 1$ odds against the null hypothesis on the $z = 4$ person.

A $z = 3.75$ threshold leaves about $200 - 1$ odds.

OK here, but not if factual rate is under 1%.
This analysis does not depend on how many of the factual positives gave positive test results.
If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case.
But the chance of getting a $z = 4$ result on the 1 brilliant player also generally goes down to 1-in-10. The confidence ratio is $60/0.1 = 600$-to-1 even so.

Sensitivity and soundness generally remain separate criteria.
This is relevant insofar as I often get a lot of 3.00–4.00 range results.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
  - So 600-1 odds against the null hypothesis on the $z = 4$ person.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
  - So 600-1 odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about 200-1 odds.
Suppose we get $z = 4$ in online chess with adult cheating rate 2%. Out of 30,000 people:

- 1 false positive result.
- 600 factual positives.
- So 600-1 odds against the null hypothesis on the $z = 4$ person.

A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%. This analysis does not depend on how many of the factual positives gave positive test results. If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also generally goes down to 1-in-10. The confidence ratio is 60/0 = 600-to-1 even so. Sensitivity and soundness generally remain separate criteria. This is relevant insofar as I often get a lot of 3.00–4.00 range results.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
  - So 600-1 odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
  - So 600-1 odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case.
Suppose we get $z = 4$ in online chess with adult cheating rate 2%. Out of 30,000 people:
- 1 false positive result.
- 600 factual positives.
- So 600-1 odds against the null hypothesis on the $z = 4$ person.

A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%.

This analysis does not depend on how many of the factual positives gave positive test results.

If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also generally goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.
Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - 1 false positive result.
  - 600 factual positives.
  - So 600-1 odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also generally goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.
- Sensitivity and soundness generally remain separate criteria.
Suppose we get $z = 4$ in online chess with adult cheating rate 2%.

Out of 30,000 people:
- 1 false positive result.
- 600 factual positives.
- So 600-1 odds against the null hypothesis on the $z = 4$ person.

A $z = 3.75$ threshold leaves about 200-1 odds. OK here, but not if factual rate is under 1%.

This analysis does not depend on how many of the factual positives gave positive test results.

If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also generally goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.

Sensitivity and soundness generally remain separate criteria.

This is relevant insofar as I often get a lot of 3.00–4.00 range results.
Time and Difficulty

Pre-Check: The “Screening” Stage

Makes a simple “box score” of agreements to the chess engine being tested and the scaled average centipawn loss from disagreements. Creates a Raw Outlier Index (ROI) from the raw metrics. ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation given one’s rating and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60. Like medical stats except indexed to common normal scale. 65 = amber alert, 70 = code orange, 75 = red. Example.

Completely data driven—no theoretical equation. Rapid and Blitz trained on in-person events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018. Does not account for the difficulty of games. That is the job of the full model.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the scaled average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the scaled average centipawn loss from disagreements.
- Creates a Raw Outlier Index (ROI) from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation given one’s rating and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation *given one’s rating* and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation *given one’s rating* and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except indexed to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation *given one’s rating* and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. **Example**.
- **Completely data driven**—no theoretical equation.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the scaled average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation *given one’s rating* and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except indexed to common normal scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- Completely data driven—no theoretical equation.
- Rapid and Blitz trained on in-person events in 2019.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation *given one’s rating* and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. **Example**.
- **Completely data driven**—no theoretical equation.
- Rapid and Blitz trained on **in-person** events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** from the raw metrics.
- ROI is on same 0-100 scale as flipping a fair coin 100 times: 50 is the expectation given one’s rating and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except indexed to common normal scale.
- 65 = amber alert, 70 = code orange, 75 = red. **Example**.
- Completely data driven—no theoretical equation.
- Rapid and Blitz trained on **in-person** events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018.
- Does not account for the difficulty of games. That is the job of the full model.
The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen. But this has perforce been post-normal science. My “back of the envelope” formula held up over two years with only one small revision for preteens. Larger 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. Velpula Sarayu.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen. But this has perforce been post-normal science.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.

But this has perforce been post-normal science.

My “back of the envelope” formula held up over two years with only one small revision for preteens.
Time and Difficulty

Rating Lag—Natural Versus Systematic

The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.

But this has perforce been post-normal science.

My “back of the envelope” formula held up over two years with only one small revision for preteens.

Larger revision in Oct. 2022 to curtail projections past Elo 2000 level.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.

But this has perforce been post-normal science.

My “back of the envelope” formula held up over two years with only one small revision for preteens.

Larger 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.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen. But this has perforce been post-normal science.

My “back of the envelope” formula held up over two years with only one small revision for preteens.

Larger 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.
The #1 scientific role I’ve played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen. But this has perforce been post-normal science.

My “back of the envelope” formula held up over two years with only one small revision for preteens.

Larger 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. Velpula Sarayu.
The article’s larger subject is a drastic proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players. (This is on top of things I’ve been telling FIDE about ratings above 2000.) My own work has been “tinged” by this issue. A natural metric apart from both my model and Sonas’s domain cross-validates his observations and arguments. I will now discuss some other applications that these solid foundations enable.
The article’s larger subject is a **drastic** proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players.
The article’s larger subject is a **drastic** proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players.

(This is on top of things I’ve been telling FIDE about ratings **above** 2000.)
The article’s larger subject is a **drastic** proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players.

(This is on top of things I’ve been telling FIDE about ratings *above* 2000.)

My own work has been “tinged” by this issue.
The article’s larger subject is a drastic proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players.

(This is on top of things I’ve been telling FIDE about ratings above 2000.)

My own work has been “tinged” by this issue.

A natural metric apart from both my model and Sonas’s domain cross-validates his observations and arguments.
The article’s larger subject is a **drastic** proposal by US statistician Jeff Sonas—long used by FIDE—to overhaul chess ratings below Elo 2000—that is, for beginning and amateur players.

(This is on top of things I’ve been telling FIDE about ratings **above** 2000.)

My own work has been “tinged” by this issue.

A natural metric **apart** from both my model and Sonas’s domain cross-validates his observations and arguments.

I will now discuss some other applications that these solid foundations enable.
Hans Niemann: Platform or Plateau?

*Ceci n’est pas un plateau*

*(celui-là, oui)*
The Gender Gap in Chess

...
The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
- How should one begin to address this question?
The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
- How should one begin to address this question?
- What data could corroborate a result—or a proposed explanation?
The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
- How should one begin to address this question?
- What data could corroborate a result—or a proposed explanation?
- Picture emerging from recent youth events...?