Cheating Detection and Cognitive Modeling at Chess

CS Distinguished Lecture, Northwestern University

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

29 May, 2024

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

A Simple Utility-Based Model

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility $\equiv$ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility $\equiv$ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility ≡ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility $\equiv$ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.
- *Alternative:* as probabilities of winning/drawing (say $p_{win} + 0.5p_{draw}$).

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility ≡ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.
- *Alternative:* as probabilities of winning/drawing (say $p_{win} + 0.5p_{draw}$).
- The model knows *nothing else** about chess.

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility ≡ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.
- *Alternative:* as probabilities of winning/drawing (say $p_{win} + 0.5p_{draw}$).
- The model knows *nothing else** about chess. No pieces, no board geometry.

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility ≡ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.
- *Alternative:* as probabilities of winning/drawing (say $p_{win} + 0.5p_{draw}$).
- The model knows *nothing else** about chess. No pieces, no board geometry.
- Only other ingredients: player skill parameters s, c, e_v (plus hyperparameters) and their correspondence to Elo chess ratings.

A Simple Utility-Based Model

- Like common econometric models under "Bounded Rationality."
- Utility ≡ values given by strong chess-playing programs (called "**engines**") to possible move choices in a series of chess positions in games by a player (or aggregate of players).
- In familiar units of *pawns* or (x100) *centipawns*.
- E.g. `+1.50` means the player to move is figuratively a pawn and a half (= `150cp`) ahead.
- *Alternative:* as probabilities of winning/drawing (say $p_{win} + 0.5p_{draw}$).
- The model knows *nothing else** about chess. No pieces, no board geometry.
- Only other ingredients: player skill parameters s, c, e_v (plus hyperparameters) and their correspondence to Elo chess ratings.
- (*The model does track how the calculated values of moves change as the engine progresses through *depths of search*.)

Elo Chess Ratings—and Why Cheat?

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

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

where $\Delta = R_P - R_O$ is the difference to your opponent's rating.

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

  where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by
$$p = \frac{1}{1 + \exp(c\Delta)}$$
where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800.

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

  where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800. Adult beginner ≈ 600, kids → 100.

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

  where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800. Adult beginner $\approx$ 600, kids $\rightarrow$ 100.
- **Stockfish 16 3544, Torch 1.0 3531, Komodo Dragon 3.3 3529.**

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800. Adult beginner $\approx$ 600, kids $\rightarrow$ 100.
- **Stockfish 16 3544, Torch 1.0 3531, Komodo Dragon 3.3 3529.**
- So computers are at "Class 15."

Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800. Adult beginner $\approx$ 600, kids $\to$ 100.
- **Stockfish 16 3544, Torch 1.0 3531, Komodo Dragon 3.3 3529.**
- So computers are at "Class 15." $\implies$ a "Moore's Law of Games."

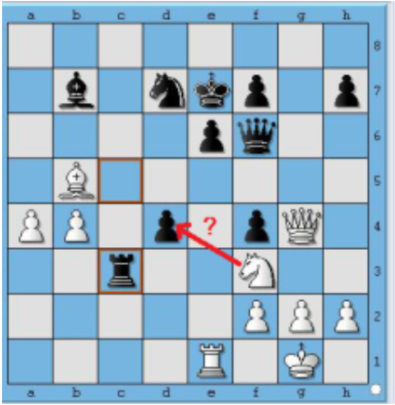
Elo Chess Ratings—and Why Cheat?

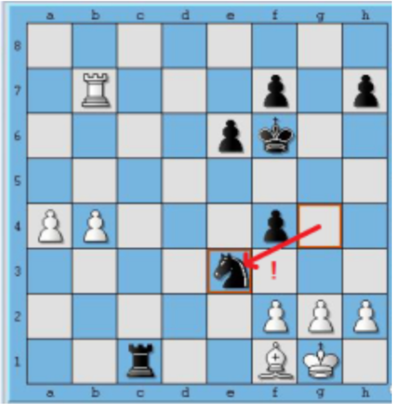
- Named for **Arpad Elo**, number R_P rates skill of player P.
- Defined by Logistic Curve: expected win % p given by

$$p = \frac{1}{1 + \exp(c\Delta)}$$

  where $\Delta = R_P - R_O$ is the difference to your opponent's rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen only player over 2800. Adult beginner $\approx$ 600, kids $\to$ 100.
- **Stockfish 16 3544, Torch 1.0 3531, Komodo Dragon 3.3 3529.**
- So computers are at "Class 15." $\implies$ a "Moore's Law of Games."
- Other Q: How do computer evaluations translate to chances of winning?

Move Utilities Example (Kramnik-Anand, 2008)





Depths... Values by Stockfish 6

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

Utility-Based Predictive Modeling

Utility-Based Predictive Modeling

- Predictive $\equiv$ model gives probabilities p_i for each option/event m_i.

Utility-Based Predictive Modeling

- Predictive ≡ model gives probabilities p_i for each option/event m_i.
- Relation to utility is usually **log-linear**:

$$\log(p_i) = \alpha + \beta u_i.$$

Utility-Based Predictive Modeling

- Predictive $\equiv$ model gives probabilities p_i for each option/event m_i.
- Relation to utility is usually **log-linear**:

$$\log(p_i) = \alpha + \beta u_i.$$

- Equivalently, if we rank options by best-first utility:

$$\log(p_1) - \log(p_i) = \beta(u_1 - u_i) \equiv \beta\delta_i.$$

Utility-Based Predictive Modeling

- Predictive $\equiv$ model gives probabilities p_i for each option/event m_i.
- Relation to utility is usually **log-linear**:

$$\log(p_i) = \alpha + \beta u_i.$$

- Equivalently, if we rank options by best-first utility:

$$\log(p_1) - \log(p_i) = \beta(u_1 - u_i) \equiv \beta\delta_i.$$

- Solved via **softmax**: $p_i = \frac{\exp(\beta\delta_i)}{\sum_{j=1}^{\ell} \exp(\beta\delta_j)}$.

Utility-Based Predictive Modeling

- Predictive ≡ model gives probabilities p_i for each option/event m_i.
- Relation to utility is usually **log-linear**:

$$\log(p_i) = \alpha + \beta u_i.$$

- Equivalently, if we rank options by best-first utility:

$$\log(p_1) - \log(p_i) = \beta(u_1 - u_i) \equiv \beta\delta_i.$$

- Solved via **softmax**: $p_i = \frac{\exp(\beta\delta_i)}{\sum_{j=1}^{\ell}\exp(\beta\delta_j)}$.
- With $\delta_1 = 0$, so that $\exp(\beta\delta_1) = 1$, this gives $p_1 = 1/\sum_{j=1}^{\ell} p_j$ and

$$p_i = p_1 \exp(-\beta\delta_i)$$

if you keep β positive.

Utility-Based Predictive Modeling

- Predictive $\equiv$ model gives probabilities p_i for each option/event m_i.
- Relation to utility is usually **log-linear**:

$$\log(p_i) = \alpha + \beta u_i.$$

- Equivalently, if we rank options by best-first utility:

$$\log(p_1) - \log(p_i) = \beta(u_1 - u_i) \equiv \beta\delta_i.$$

- Solved via **softmax**: $p_i = \frac{\exp(\beta\delta_i)}{\sum_{j=1}^{\ell} \exp(\beta\delta_j)}$.
- With $\delta_1 = 0$, so that $\exp(\beta\delta_1) = 1$, this gives $p_1 = 1/\sum_{j=1}^{\ell} p_j$ and

$$p_i = p_1 \exp(-\beta\delta_i)$$

if you keep β positive. Probabilities are **multiples** of p_1.

Loglog-Linear Model

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

Equivalently,

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

Loglog-Linear Model

$$\log\log\left(\frac{1}{p_i}\right) - \log\log\left(\frac{1}{p_1}\right) = \beta\delta_i.$$

Equivalently,

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

This gives

$$p_i = p_1^{\exp(\beta u_i)},$$

so that probabilities are represented as **powers** of p_1.

Loglog-Linear Model

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

Equivalently,

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

This gives

$$p_i = p_1^{\exp(\beta u_i)},$$

so that probabilities are represented as **powers** of p_1.

A rare bird?

Loglog-Linear Model

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

Equivalently,

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

This gives

$$p_i = p_1^{\exp(\beta u_i)},$$

so that probabilities are represented as **powers** of p_1.

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

Parameters and Nonlinearity

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i,

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i, $\implies$ better **strategic** ability to perceive small advantages.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i, $\implies$ better **strategic** ability to perceive small advantages. Like Anatoly Karpov.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i, $\implies$ better **strategic** ability to perceive small advantages. Like Anatoly Karpov.
- Higher (=better) **consistency** c drives down p_i for moves of large δ_i,

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i, $\implies$ better **strategic** ability to perceive small advantages. Like Anatoly Karpov.
- Higher (=better) **consistency** c drives down p_i for moves of large δ_i, ability to survive **tactical** minefields.

Parameters and Nonlinearity

- Note β cancels the centipawn units of δ_i, so we write $\frac{\delta_i}{s}$ instead.
- Since $\frac{\delta_i}{s}$ is dimensionless, can raise to any power c.
- Basic log-linear model becomes: $p_i = p_1 \cdot \exp(-\left(\frac{\delta_i}{s}\right)^c)$.
- Double-log model becomes: $p_i = p_1^{\exp(\left(\frac{\delta_i}{s}\right)^c)}$.

Intuition either way:

- Lower (=better) **sensitivity** s magnifies effect of small δ_i, $\implies$ better **strategic** ability to perceive small advantages. Like Anatoly Karpov.
- Higher (=better) **consistency** c drives down p_i for moves of large δ_i, ability to survive **tactical** minefields. Like Mikhail Tal.

Karpov & Tal at Montreal "Tourney of Stars" 1979

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s = \mathbf{0.01558}$, $c = \mathbf{0.30702}$.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558,** $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623,** $c =$ **0.36474**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558,** $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623,** $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558,** $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623,** $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.
- These are my **Intrinsic Performance Ratings** (**IPRs**).

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558**, $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623**, $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.
- These are my **Intrinsic Performance Ratings** (**IPRs**).
- Whole tourney IPR is (only!) **2575 +- 50**.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558**, $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623**, $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.
- These are my **Intrinsic Performance Ratings** (**IPRs**).
- Whole tourney IPR is (only!) **2575 +- 50**. (With $s = 0.04121$, $c = 0.38525$.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558**, $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623**, $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.
- These are my **Intrinsic Performance Ratings** (**IPRs**).
- Whole tourney IPR is (only!) **2575 +- 50**. (With $s = 0.04121$, $c = 0.38525$.
- Average Elo of players, **2621**, is within error bars.

Karpov & Tal at Montreal "Tourney of Stars" 1979

- Tied for first with 12/18 in star-studded double round-robin.
- Karpov was rated **2705**, Tal only **2615**.
- Karpov (per SF11): $s =$ **0.01558,** $c =$ **0.30702**.
- Tal (per SF11): $s =$ **0.02623,** $c =$ **0.36474**.
- Trained correspondence to Elo rating gives Karpov **2625 +- 155**, Tal **2730 +- 185**.
- These are my **Intrinsic Performance Ratings** (**IPRs**).
- Whole tourney IPR is (only!) **2575 +- 50**. (With $s = 0.04121$, $c = 0.38525$.
- Average Elo of players, **2621**, is within error bars. Surprise is that the IPR is not near 2700s range.

Test Quantities and Parameter Fitting

Over T-many game turns t by a player (or players), solve to make the following two test quantities into **unbiased estimators**:

Test Quantities and Parameter Fitting

Over T-many game turns t by a player (or players), solve to make the following two test quantities into **unbiased estimators**:

- **T1-Match**: Make the actual number t_a^1 of agreements with the engine equal

$$t_{proj}^1 = \sum_{t=1}^{T} p_{1,t}.$$

Test Quantities and Parameter Fitting

Over T-many game turns t by a player (or players), solve to make the following two test quantities into **unbiased estimators**:

- **T1-Match**: Make the actual number t_a^1 of agreements with the engine equal

$$t_{proj}^1 = \sum_{t=1}^{T} p_{1,t}.$$

- **ASD**: Make the *scaled* "average centipawn loss" asd_a of a player's moves $m_{i_t,t}$—as judged by the testing engine—equal

$$asd_{proj} = \sum_{t=1}^{T} \sum_{i=1}^{\ell} p_{i,t}\delta_{i,t}.$$

Alternative fitting methods include maximum-likelihood estimation, equivalently, minimzing $\sum_{t=1}^{T} \log(\frac{1}{p_{i_t,t}})$.

Other Quantities of Interest

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match."

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.
- **T3, M3, etc.** "T3-match" much-discussed cheating metric.

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.
- **T3, M3, etc.** "T3-match" much-discussed cheating metric.
- **Error100**: Mistakes m_i with $\delta_i \geq 100$ (i.e., one pawn).

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.
- **T3, M3, etc.** "T3-match" much-discussed cheating metric.
- **Error100**: Mistakes m_i with $\delta_i \geq 100$ (i.e., one pawn).
- **Error200**: Moves m_i with $\delta_i \geq 200$, "game-losing blunders."

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.
- **T3, M3, etc.** "T3-match" much-discussed cheating metric.
- **Error100**: Mistakes m_i with $\delta_i \geq 100$ (i.e., one pawn).
- **Error200**: Moves m_i with $\delta_i \geq 200$, "game-losing blunders."
- **Delta**(u, v): moves with $u \leq \delta_i \leq v$, "small slips."

Other Quantities of Interest

- **EV-Match**: About 8–10% of positions have multiple optimal moves. Include them all as a "match."
- **T2-Match**: Include the second-best move as a "match." (Unless it is a blunder...)
- **M2**: $p_{2,t}$ vs. actual frequency of playing second-best move.
- **T3, M3, etc.** "T3-match" much-discussed cheating metric.
- **Error100**: Mistakes m_i with $\delta_i \geq 100$ (i.e., one pawn).
- **Error200**: Moves m_i with $\delta_i \geq 200$, "game-losing blunders."
- **Delta**(u, v): moves with $u \leq \delta_i \leq v$, "small slips."
- Captures, advancing vs. retreating moves, moves with Knights or other specific pieces...

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.**

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.** Makes model unstable. Similar issue.

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.** Makes model unstable. Similar issue.
- Coped with by replacing h by the parameter e_v, which leverages "swing" of moves whose highest-depth value is equal-optimal, so as to fit **EV** as a third unbiased estimator.

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.** Makes model unstable. Similar issue.
- Coped with by replacing h by the parameter e_v, which leverages "swing" of moves whose highest-depth value is equal-optimal, so as to fit **EV** as a third unbiased estimator.
- This enables deploying **EV** as a z-test alongside **T1** and **ASD**.

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.** Makes model unstable. Similar issue.
- Coped with by replacing h by the parameter e_v, which leverages "swing" of moves whose highest-depth value is equal-optimal, so as to fit **EV** as a third unbiased estimator.
- This enables deploying **EV** as a z-test alongside **T1** and **ASD**.
- Idea of ρ_i still impacts r_i and hence s and c.

Improving Predictivity

Original Idea (2015–2017): Add a term ρ_i for "perceived" (change in) value over lower depths of search. Higher for "trappy" moves. Multiply by third parameter h:

$$r_i = \left(\frac{\delta_i + h\rho_i}{s}\right)^c.$$

- **Problem: Observed $h \gg 1$.** Makes model unstable. Similar issue.
- Coped with by replacing h by the parameter e_v, which leverages "swing" of moves whose highest-depth value is equal-optimal, so as to fit **EV** as a third unbiased estimator.
- This enables deploying **EV** as a z-test alongside **T1** and **ASD**.
- Idea of ρ_i still impacts r_i and hence s and c.
- Enables projecting some inferior move as more likely than m_1 in about 15% of positions, improving the "prediction hit" rate by 2–3 percentage points.

Demonstration: 2024 FIDE Candidates Tournaments

(show)

Happy Birthday 29 May to the winners, D. Gukesh and Zhongyi Tan!

Basic Model Sanity Facts

Whereas the fitted log-linear model *grossly underestimates* **M2** and **M3**, the fitted double-log model underestimates them (hence also **T2** and **T3**) only slightly. Moreover:

For each other metric μ, the "ersatz z-test"

$$z_\mu = \frac{\mu_a - \mu_{proj}}{\sigma_\mu}$$

is tolerably close to Gaussian normal $\mathcal{N}(0,1)$ and with considerable independence of other $z_{\mu'}$. This is so both after fitting and under the rating-based testing procedure.

The main quantities z_{T1}, z_{ASD}, and z_{EV} are expressly **adjusted** to conform to the (upper arm of the) bell curve in myriad **randomized resampling** trials over (parts of) the training sets.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly “chessy” enough.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*. (Voiceover: They're not.)

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*. (Voiceover: They're not.)
- But it is a *sparse, nearest-neighbor dependence*, hence approximable by scalar means without having to model big covariance matrices.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*. (Voiceover: They're not.)
- But it is a *sparse, nearest-neighbor dependence*, hence approximable by scalar means without having to model big covariance matrices.
- Gets done empirically via said resampling trials.

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*. (Voiceover: They're not.)
- But it is a *sparse, nearest-neighbor dependence*, hence approximable by scalar means without having to model big covariance matrices.
- Gets done empirically via said resampling trials.
- That ensures **safety** (against false positives).

Cheating Test Sanity and Sensitivity

Say we test a player on $T = 200$ relevant moves across 9 games.

- Because **T1**, **EV**, and **ASD** are aggregate quantities—averages—the **Central Limit Theorem** takes hold...
- ...despite the 200 positions not drawing from the same distribution of plausible moves.
- The distributions are (evidently) similarly "chessy" enough.
- Simple computation of the projected σ_{T1}, σ_{EV}, and σ_{ASD} presumes that the positions-and-their-choices are *independent*. (Voiceover: They're not.)
- But it is a *sparse, nearest-neighbor dependence*, hence approximable by scalar means without having to model big covariance matrices.
- Gets done empirically via said resampling trials.
- That ensures **safety** (against false positives). How about **sensitivity** (avoiding false negatves)?

Cognitive Concepts and Conceits

Many results in cognitive decision making come from studies that

1. are well-targeted to the concept and hypothesis, but

Cognitive Concepts and Conceits

Many results in cognitive decision making come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...

Cognitive Concepts and Conceits

Many results in cognitive decision making come from studies that

1. are well-targeted to the concept and hypothesis, but
2. have under 100 test subjects...
3. ...under simulated conditions...

Cognitive Concepts and Conceits

Many results in cognitive decision making come from studies that

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

Cognitive Concepts and Conceits

Many results in cognitive decision making come from studies that

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

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

Some Accompanying Stances

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 **satisfice**.

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:

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

1. 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

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.

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).

Player Development

Player Development

5. Rating Inflation? Deflation?

Player Development

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

Player Development

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

Player Development

5. 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."

Player Development

5. 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?

Player Development

5. 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

5. 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?
6. Growth Curves of Improving (Young) Players.

Player Development

5. 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?
6. Growth Curves of Improving (Young) Players.
7. Relationship of Quality to Thinking Time Budget. (show graph) (or this)

7. (New) Time Management

The Women's Candidates used the FIDE Standard time control:

- 90 minutes at the start.

7. (New) Time Management

The Women's Candidates used the FIDE Standard time control:

- 90 minutes at the start.
- 30 seconds **increment** starting from the frst move.

7. (New) Time Management

The Women's Candidates used the FIDE Standard time control:

- 90 minutes at the start.
- 30 seconds **increment** starting from the frst move.
- 30 minute "lump sum" added after turn 40.

Gives 110 minutes to the turn 40 "time control" and 150 minutes to turn 60.

The Open (Men's) section gave 120 minutes at the start, with 30 minute lump sum after turn 40, **but** 30 seconds increment only after turn 40. Thus the moves up to turn 40 were "classic time pressure" without increment. (Gives only 160 minutes to turn 60.)

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.
- Spending 15 minutes or more gives even worse performance.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.
- Spending 15 minutes or more gives even worse performance.
- **Is Thinking Bad For You?**

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.
- Spending 15 minutes or more gives even worse performance.
- **Is Thinking Bad For You?**
- Similar phenomena observed in blitz chess by Ashton Anderson (UT), Jon Kleinberg (Cornell), and others in and apart from their group, from giant corpus of online games.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.
- Spending 15 minutes or more gives even worse performance.
- **Is Thinking Bad For You?**
- Similar phenomena observed in blitz chess by Ashton Anderson (UT), Jon Kleinberg (Cornell), and others in and apart from their group, from giant corpus of online games.
- If we include **having little time left** into the predicate—average before turn 40 or overall—then results are closer to expectation.

Candidates for Shock?

Let's first combine the sections and look at positions where players spent a lot or a little time, irrespective of time pressure.

- Combined, they played close to their **2627** rating average.
- Predicated on making their move within **5 seconds** they played...**well over 3000 level**.
- Predicated on spending at least 10 minutes on a move, they played...**only about 2200 level**.
- Spending 15 minutes or more gives even worse performance.
- **Is Thinking Bad For You?**
- Similar phenomena observed in blitz chess by Ashton Anderson (UT), Jon Kleinberg (Cornell), and others in and apart from their group, from giant corpus of online games.
- If we include **having little time left** into the predicate—average before turn 40 or overall—then results are closer to expectation.
- (From my recent graduate seminar. Q&A phase can begin here.)

8. How to Measure “Difficulty”?

8. How to Measure "Difficulty"?

- Does it equal "Hazard"—meaning the expected loss of value (and of win/draw probability) from the choice of move?

8. How to Measure "Difficulty"?

- Does it equal "Hazard"—meaning the expected loss of value (and of win/draw probability) from the choice of move?
- Or does it have more to do with the chance of finding an optimal move?

8. How to Measure "Difficulty"?

- Does it equal "Hazard"—meaning the expected loss of value (and of win/draw probability) from the choice of move?
- Or does it have more to do with the chance of finding an optimal move?
- Correspondence to Multiple-Choice Tests.

8. How to Measure "Difficulty"?

- Does it equal "Hazard"—meaning the expected loss of value (and of win/draw probability) from the choice of move?
- Or does it have more to do with the chance of finding an optimal move?
- Correspondence to Multiple-Choice Tests.
- The "Solitaire Chess" feature by Bruce Pandolfini gives partial credits for reasonable moves.

Entropy and Difficulty

Is Hazard maximized when

Entropy and Difficulty

Is Hazard maximized when

- there are many tempting, somewhat-inferior moves? (High entopy)

Entropy and Difficulty

Is Hazard maximized when

- there are many tempting, somewhat-inferior moves? (High entopy)
- Or when all moves except one are tangibly inferior? (Lower entropy)

Entropy and Difficulty

Is Hazard maximized when

- there are many tempting, somewhat-inferior moves? (High entopy)
- Or when all moves except one are tangibly inferior? (Lower entropy)
- Results from my seminar show that difficulty goes with entropy more than previously expected.

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

- **T1** match: No—it will show lower match rates in high-entropy positions.

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

- **T1** match: No—it will show lower match rates in high-entropy positions.
- ASD metric—?

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

- **T1** match: No—it will show lower match rates in high-entropy positions.
- ASD metric—?
- IPR metric—??

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

- **T1** match: No—it will show lower match rates in high-entropy positions.
- ASD metric—?
- IPR metric—?? By intent, this *should* give signal consistency.

9. Signal Consistency

Suppose we know an overall Elo skill level E for a set of players in advance. On (which) subsets of the data should we expect a metric μ to give consistent readings in the vicinity of E?

- **T1** match: No—it will show lower match rates in high-entropy positions.
- ASD metric—?
- IPR metric—?? By intent, this *should* give signal consistency.
- Reasonable on, say, positions with +1.00 or more advantage, versus positions with -1.00 or worse disadvantage, versus evenly balanced positions.

Examination Grading Analogy

I typically design exams to have about

- 20% A-level questions (and points)
- 30% B-level,
- 30% C-level, and
- 20% D-level, with 90% the target for an A grade.

Means that getting 60% on the A-level questions is reasonably on-track for an A, even though 60% by itself is a "C signal."

Examination Grading Analogy

I typically design exams to have about

- 20% A-level questions (and points)
- 30% B-level,
- 30% C-level, and
- 20% D-level, with 90% the target for an A grade.

Means that getting 60% on the A-level questions is reasonably on-track for an A, even though 60% by itself is a “C signal.”

Should we use metrics that would say “A” even on the difficult questions by themselves, rather than rely on the exam being overall farly designed?

Examination Grading Analogy

I typically design exams to have about

- 20% A-level questions (and points)
- 30% B-level,
- 30% C-level, and
- 20% D-level, with 90% the target for an A grade.

Means that getting 60% on the A-level questions is reasonably on-track for an A, even though 60% by itself is a "C signal."

Should we use metrics that would say "A" even on the difficult questions by themselves, rather than rely on the exam being overall farly designed? Matters for *adaptive-difficulty* automated exams, which grade you by finding the level at which you score 50% (or 75% or etc). (IRT theory again).

Conclusions and Future Work

Q&A and Thanks

Cancer and Covid (= in-person and online chess)

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...

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?*

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.

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 $\approx$ 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.
- 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 $\approx$ 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

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: **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.

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: **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.

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.

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.

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

Covid in Non-Surge and Surge Times

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

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.

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.

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*:

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.

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.

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.

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.*

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.

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?

Back to Chess...

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.

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%.

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.

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.

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.*

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.*
- This is relevant insofar as I often get a lot of 3.00–4.00 range results.

Internal and External Confidence

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)
- $\implies$ covariance matrix is banded, hence approximable by scalars.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)
- $\implies$ covariance matrix is banded, hence approximable by scalars.
- Could treat as a "reduced-entropy" sample size $T' < T$.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)
- $\implies$ covariance matrix is banded, hence approximable by scalars.
- Could treat as a "reduced-entropy" sample size $T' < T$.
- What I actually do is adjust σ up to σ'_E with dependence on Elo rating E determined by millions of **randomized resampling** trials from the training sets.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)
- $\implies$ covariance matrix is banded, hence approximable by scalars.
- Could treat as a "reduced-entropy" sample size $T' < T$.
- What I actually do is adjust σ up to σ'_E with dependence on Elo rating E determined by millions of **randomized resampling** trials from the training sets.
- With this patched, justified in saying the model paints chess moves on a 1,000-sided die and *simply rolls it*.

Internal and External Confidence

- Projections also automatically give additive variance, hence σ and confidence intervals, **if** we assume turn decisions are *independent*.
- [VOICEOVER: **They're not.**]
- But it's a *sparse dependence* on neighboring moves. (Not across games—common "opening book" is removed from the sample.)
- $\implies$ covariance matrix is banded, hence approximable by scalars.
- Could treat as a "reduced-entropy" sample size $T' < T$.
- What I actually do is adjust σ up to σ'_E with dependence on Elo rating E determined by millions of **randomized resampling** trials from the training sets.
- With this patched, justified in saying the model paints chess moves on a 1,000-sided die and *simply rolls it.* $\implies$ multinomial Bernoulli trials.

Pre-Check: The "Screening" Stage

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)** on the same 0-100 scale as flipping a fair coin 100 times.

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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 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.

Z-Scores and Cheating Tests

For the aggregate quantities, the Central Limit Theorem in practice allows treating

$$z' = \frac{(\text{actual}) - (\text{predicted})}{\sigma'}$$

as a ***z*-score** (after adjustment).

Evaluation Criteria:

Z-Scores and Cheating Tests

For the aggregate quantities, the Central Limit Theorem in practice allows treating

$$z' = \frac{\text{(actual)} - \text{(predicted)}}{\sigma'}$$

as a **z-score** (after adjustment).

Evaluation Criteria:

- **Safety**: Over fair=playing populations, $z' \sim$ bell curve.

Z-Scores and Cheating Tests

For the aggregate quantities, the Central Limit Theorem in practice allows treating

$$z' = \frac{(\text{actual}) - (\text{predicted})}{\sigma'}$$

as a ***z*-score** (after adjustment).

Evaluation Criteria:

- **Safety**: Over fair=playing populations, $z' \sim$ bell curve.
- **Sensitivity**: Factual cheaters yield "high enough" z'.

From this point on, let's suppose my model has these properties. What about interpreting the results?

Suppose We Get $z = 3.54$

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000.

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives "raw face-value odds" of "5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is
  - 1-in-5,000 to 1-in-10,000 for in-person chess.

Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives “raw face-value odds” of “5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is
    - 1-in-5,000 to 1-in-10,000 for in-person chess.
    - 1-in-50 (greater for kids) to 1-in-200 for online chess.

Suppose We Get $z = 3.54$

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

Suppose We Get $z = 3.54$

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

Suppose We Get $z = 3.54$

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

Suppose We Get $z = 3.54$

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

Suppose We Get $z = 3.54$

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

Are these considerations orthogonal, or do they align?

Fraught Issue #1

What should be the target confidence?

Fraught Issue #1

What should be the target confidence?

1. Proof beyond reasonable doubt?

Fraught Issue #1

What should be the target confidence?

1. Proof beyond reasonable doubt?
2. **“Comfortable satisfaction”**

Fraught Issue #1

What should be the target confidence?

1. Proof beyond reasonable doubt?
2. **"Comfortable satisfaction"**
3. **"Balance of Probability"**

CAS Lausanne recognizes all three, but inclines toward 2.

- Still doesn't specify a corresponding confidence target.

Fraught Issue #1

What should be the target confidence?

1. Proof beyond reasonable doubt?
2. **"Comfortable satisfaction"**
3. **"Balance of Probability"**

CAS Lausanne recognizes all three, but inclines toward 2.

- Still doesn't specify a corresponding confidence target.
- Science, of course, demands criterion 1.

Fraught Issue #2: Confidence For Chess

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a “Florida” every 4 cycles, because returns often blast past that instantly.

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a "Florida" every 4 cycles, because returns often blast past that instantly.
- So maybe truer chess analogue is 1-in-500 error.

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a "Florida" every 4 cycles, because returns often blast past that instantly.
- So maybe truer chess analogue is 1-in-500 error.
- Judge by **"Countenanced Error Rate Per Year."**

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a "Florida" every 4 cycles, because returns often blast past that instantly.
- So maybe truer chess analogue is 1-in-500 error.
- Judge by **"Countenanced Error Rate Per Year."**
- E.g. if 10 cases per year reach judgment stage, and you can tolerate 1 error per 20 years, then 99.5

Fraught Issue #2: Confidence For Chess

- **I** interpret the range of comfortable satisfaction as **99–99.9%** final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a "Florida" every 4 cycles, because returns often blast past that instantly.
- So maybe truer chess analogue is 1-in-500 error.
- Judge by **"Countenanced Error Rate Per Year."**
- E.g. if 10 cases per year reach judgment stage, and you can tolerate 1 error per 20 years, then 99.5
- But online chess has 10,000+ cases per year...

Issue # 3: Accounting "Look Elsewhere "

Issue # 3: Accounting "Look Elsewhere "

- Approximately 100,000 players-in-event per year among "notable" events.

Issue # 3: Accounting “Look Elsewhere ”

- Approximately 100,000 players-in-event per year among “notable” events.
  - notable $\equiv$ some or all gamescores preserved.

Issue # 3: Accounting "Look Elsewhere "

- Approximately 100,000 players-in-event per year among "notable" events.
  - notable $\equiv$ some or all gamescores preserved.
- A highly computerlike game is a "shiny marble"—players do notice.

Issue # 3: Accounting "Look Elsewhere "

- Approximately 100,000 players-in-event per year among "notable" events.
  - notable ≡ some or all gamescores preserved.
- A highly computerlike game is a "shiny marble"—players do notice.
- Accounted over a year, suggests to divide odds by 100,000.

Issue # 3: Accounting "Look Elsewhere "

- Approximately 100,000 players-in-event per year among "notable" events.
  - notable ≡ some or all gamescores preserved.
- A highly computerlike game is a "shiny marble"—players do notice.
- Accounted over a year, suggests to divide odds by 100,000.
  - 4.75 sigma $\longrightarrow$ only 90% confidence.
  - 5.00 sigma $\longrightarrow$ 1-in-35 error.

Issue # 3: Accounting “Look Elsewhere ”

- Approximately 100,000 players-in-event per year among “notable” events.
    - notable $\equiv$ some or all gamescores preserved.
- A highly computerlike game is a “shiny marble”—players do notice.
- Accounted over a year, suggests to divide odds by 100,000.
    - 4.75 sigma $\longrightarrow$ only 90% confidence.
    - 5.00 sigma $\longrightarrow$ 1-in-35 error.
- Sounds like 1-in-35 error is still too high based on confidence target.

Issue # 3: Accounting "Look Elsewhere "

- Approximately 100,000 players-in-event per year among "notable" events.
  - notable ≡ some or all gamescores preserved.
- A highly computerlike game is a "shiny marble"—players do notice.
- Accounted over a year, suggests to divide odds by 100,000.
  - 4.75 sigma ⟶ only 90% confidence.
  - 5.00 sigma ⟶ 1-in-35 error.
- Sounds like 1-in-35 error is still too high based on confidence target.
- But reckon against time-scale of actual cases and tolerated error rate.

.

Doomsday to the Rescue?

Why stop at a year? Why not consider “look elsewhere” over an entire 50-year span?

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)

Doomsday to the Rescue?

Why stop at a year? Why not consider “look elsewhere” over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)

Doomsday to the Rescue?

Why stop at a year? Why not consider “look elsewhere” over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)
- Better argument?: Balance against the arrival rate of real cases.

Doomsday to the Rescue?

Why stop at a year? Why not consider “look elsewhere” over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)
- Better argument?: Balance against the arrival rate of real cases.
- Aligns with Bayesian prior on average, but should allow for variance in the rate.

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)
- Better argument?: Balance against the arrival rate of real cases.
- Aligns with Bayesian prior on average, but should allow for variance in the rate.
- Figure discount by 25,000 to 50,000.

Doomsday to the Rescue?

Why stop at a year? Why not consider "look elsewhere" over an entire 50-year span?

- IMHO, the notorious **Doomsday Argument** kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event **this (early) year?**
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)
- Better argument?: Balance against the arrival rate of real cases.
- Aligns with Bayesian prior on average, but should allow for variance in the rate.
- Figure discount by 25,000 to 50,000. Then 5-sigma is OK.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.
- Qualifying events for championships.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.
- Qualifying events for championships.
- Major international Opens.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.
- Qualifying events for championships.
- Major international Opens.
- The Carlsen Online Chess Tour.

Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.
- Qualifying events for championships.
- Major international Opens.
- The Carlsen Online Chess Tour.
- Chess.com "Titled Tuesdays" ...

The combination of the online 100-1 prior and marquee online events amps up the calculus.

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a “marked man”?

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a "marked man"? Even granting he's never cheated at in-person chess?

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a "marked man"? Even granting he's never cheated at in-person chess?

- Niemann plays ≈ 25 events per year.

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a “marked man”? Even granting he’s never cheated at in-person chess?

- Niemann plays ≈ 25 events per year.
- Like giving drug test to same athlete 25x.

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a "marked man"? Even granting he's never cheated at in-person chess?

- Niemann plays ≈ 25 events per year.
- Like giving drug test to same athlete 25x.
- But what about a player wearing a heavy winter overcoat in hot weather?

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a “marked man”? Even granting he’s never cheated at in-person chess?

- Niemann plays ≈ 25 events per year.
- Like giving drug test to same athlete 25x.
- But what about a player wearing a heavy winter overcoat in hot weather?
- Or a player wearing neon-green sneakers??

Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a “marked man”? Even granting he’s never cheated at in-person chess?

- Niemann plays $\approx$ 25 events per year.
- Like giving drug test to same athlete 25x.
- But what about a player wearing a heavy winter overcoat in hot weather?
- Or a player wearing neon-green sneakers??
- Yet another separate matter from the Bayesian prior.

Super-Fraught Issue #6: Multi-Testing Samples

Super-Fraught Issue #6: Multi-Testing Samples

- Includes **Cherry-Picking** and other forms of ***p*-hacking**.

Super-Fraught Issue #6: Multi-Testing Samples

- Includes **Cherry-Picking** and other forms of p-**hacking**.
- What if a player seems to have cheated only in games 5–8 of a nine-game Open?

Super-Fraught Issue #6: Multi-Testing Samples

- Includes **Cherry-Picking** and other forms of ***p*-hacking**.
- What if a player seems to have cheated only in games 5–8 of a nine-game Open?
- Or maybe games 4–6 and 8–9?

Super-Fraught Issue #6: Multi-Testing Samples

- Includes **Cherry-Picking** and other forms of **p-hacking**.
- What if a player seems to have cheated only in games 5–8 of a nine-game Open?
- Or maybe games 4–6 and 8–9?
- Proper domain of Bonferroni Correction if it doesn't wipe out significance altogether.

Super-Fraught Issue #6: Multi-Testing Samples

- Includes **Cherry-Picking** and other forms of p-**hacking**.
- What if a player seems to have cheated only in games 5–8 of a nine-game Open?
- Or maybe games 4–6 and 8–9?
- Proper domain of Bonferroni Correction if it doesn't wipe out significance altogether.
- Well, z-hacking/p-hacking is a huge area...

Issue #7: Results on Aggregates of Players

Issue #7: Results on Aggregates of Players

- What if you get $z = 3.54$ on three different players in a 500-player Open?

Issue #7: Results on Aggregates of Players

- What if you get $z = 3.54$ on three different players in a 500-player Open?
- Not enough to convict any one player.

Issue #7: Results on Aggregates of Players

- What if you get $z = 3.54$ on three different players in a 500-player Open?
- Not enough to convict any one player.
- But odds against all being fair can be estimated by aggregating z-scores, presuming (under the null hypothesis of fair play) that the players' actions are independent:

$$z = \frac{z_1 + z_2 + z_3}{\sqrt{3}} \approx 6.13$$

Issue #7: Results on Aggregates of Players

- What if you get $z = 3.54$ on three different players in a 500-player Open?
- Not enough to convict any one player.
- But odds against all being fair can be estimated by aggregating z-scores, presuming (under the null hypothesis of fair play) that the players' actions are independent:

$$z = \frac{z_1 + z_2 + z_3}{\sqrt{3}} \approx 6.13 \text{ Billion-to-one}$$

Applying "Look-Elsewhere" still leaves astronomical confidence that *some* cheating occurred.

Issue #7: Results on Aggregates of Players

- What if you get $z = 3.54$ on three different players in a 500-player Open?
- Not enough to convict any one player.
- But odds against all being fair can be estimated by aggregating z-scores, presuming (under the null hypothesis of fair play) that the players' actions are independent:

$$z = \frac{z_1 + z_2 + z_3}{\sqrt{3}} \approx 6.13 \text{ Billion-to-one}$$

Applying "Look-Elsewhere" still leaves astronomical confidence that *some* cheating occurred. Still leaves the question of who.

Issue #8: Scaling of Estimation Error

Issue #8: Scaling of Estimation Error

- My formulas—“screening” as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:

Issue #8: Scaling of Estimation Error

- My formulas—"screening" as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robins;
  - 12–24 game championship matches.

Issue #8: Scaling of Estimation Error

- My formulas—"screening" as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robins;
  - 12–24 game championship matches.
- But how about 300+ games played in "Titled Tuesdays" over a half-year span?

Issue #8: Scaling of Estimation Error

- My formulas—"screening" as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robins;
  - 12–24 game championship matches.
- But how about 300+ games played in "Titled Tuesdays" over a half-year span?
- Skew from rating estimation error scales *linearly* as $\Omega(n)$.

Issue #8: Scaling of Estimation Error

- My formulas—"screening" as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robins;
  - 12–24 game championship matches.
- But how about 300+ games played in "Titled Tuesdays" over a half-year span?
- Skew from rating estimation error scales *linearly* as $\Omega(n)$.
- Overflows the $O(\sqrt{n})$ levees...

Issue #8: Scaling of Estimation Error

- My formulas—"screening" as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size n of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robins;
  - 12–24 game championship matches.
- But how about 300+ games played in "Titled Tuesdays" over a half-year span?
- Skew from rating estimation error scales *linearly* as $\Omega(n)$.
- Overflows the $O(\sqrt{n})$ levees... Validation by myriad resampling trials done on $n = 4, 9, 16$.

Issue #9: Biased Inputs

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. “Pandemic Lag” article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.
- As in issue 8, rating estimation bias skews linearly.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.
- As in issue 8, rating estimation bias skews linearly.
- My model has enough cross-checks to detect and correct the bias—

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.
- As in issue 8, rating estimation bias skews linearly.
- My model has enough cross-checks to detect and correct the bias—mainly need only assume not everyone is cheating.

Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. "Pandemic Lag" article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.
- As in issue 8, rating estimation bias skews linearly.
- My model has enough cross-checks to detect and correct the bias—mainly need only assume not everyone is cheating. No "interstellar dust" issue.

Going Post-Normal

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
    - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.
  -

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.
  -
  - Formula for teenagers (with 15 multiplier) otherwise *unchanged*.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.
  -
  - Formula for teenagers (with 15 multiplier) otherwise *unchanged*.
- Adjusted players are often over half the entrants in large Opens.

Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year age exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the *same* formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.
  -
  - Formula for teenagers (with 15 multiplier) otherwise *unchanged*.
- Adjusted players are often over half the entrants in large Opens.
- Basically running a more accurate rating system from the back of an envelope.

Post-Normal II: Time Dependence

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.
- Not enough reliable training data for (in-person) fast chess across skill levels.

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.
- Not enough reliable training data for (in-person) fast chess across skill levels.
- Panoply of different speeds anyway: $\tau =$ time you can use to play 60 moves.

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.
- Not enough reliable training data for (in-person) fast chess across skill levels.
- Panoply of different speeds anyway: τ = time you can use to play 60 moves.
- FIDE standard slow chess gives $\tau = 150$ minutes.

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.
- Not enough reliable training data for (in-person) fast chess across skill levels.
- Panoply of different speeds anyway: $\tau =$ time you can use to play 60 moves.
- FIDE standard slow chess gives $\tau = 150$ minutes.
- Postulate: Elo reduction $R_E(\tau)$ if largely independent of the player's Elo rating E.

Post-Normal II: Time Dependence

- The pandemic drove major tournaments online—where chess is played faster.
- Not enough reliable training data for (in-person) fast chess across skill levels.
- Panoply of different speeds anyway: $\tau =$ time you can use to play 60 moves.
- FIDE standard slow chess gives $\tau = 150$ minutes.
- Postulate: Elo reduction $R_E(\tau)$ if largely independent of the player's Elo rating E.
- Reasonable *a-priori* since chess rating system is designed for additive invariance: only the difference in ratings to the opponent matters for predictions.

Laws of Time and Difficulty

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
    - Rating estimate of $\tau = 0$, i.e., of completely random chess.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
    - Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
  - Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.
  - Aitken Extrapolation.

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
    - Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.
    - Aitken Extrapolation.
- Lo and behold—the two methods agree!

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
  - Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.
  - Aitken Extrapolation.
- Lo and behold—the two methods agree!
- Is the resuting "Rating Time Curve" thereby a natural law?

Laws of Time and Difficulty

- Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.
- Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in τ.
- Gives four unknowns to fit, but only three equations. Try getting fourth from:
  - Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.
  - Aitken Extrapolation.
- Lo and behold—the two methods agree!
- Is the resuting “Rating Time Curve” thereby a natural law?
- Does this make *time* fungible with *difficulty*, the latter as modeled by Item Response Theory?

Stance on Data Science

Stance on Data Science

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

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- Concern: Data modelers in less-extreme settings **satisfice**.

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- 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.

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- 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. (Compare what Scott Aaronson calls the Meatspace.)

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- 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. (Compare what Scott Aaronson calls the Meatspace.)
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- 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. (Compare what Scott Aaronson calls the Meatspace.)
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.
- Here is a way of phrasing the question that comes from this stance:

Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- 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. (Compare what Scott Aaronson calls the Meatspace.)
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.
- Here is a way of phrasing the question that comes from this stance:

When is it important that our models include gravity?

Q & A

And Thanks.