What Laws Act on the Mind?

Large data, regularities, and illusions

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

RKMVERI, 5 Feb. 2019

[1] Joint work with Tamal Tanu Biswas and with grateful acknowledgment to UB's Center for Computational Research (CCR)

Competitive Chess

- Burgeoning popularity and participation despite computers having dethroned human champions 22 years ago.
- India has 59 Grandmasters, including several of the youngest ones. . . 59 more than 40 years ago. (The first was V. Anand in 1988.) Bangladesh has 5. BAN championships now prominent.
- Many schools have adopted programmes in chess.
- Over this decade, many more games by amateur players have been preserved and archived in publicly available game collections.
- In 2018, I took data from 10.6 million positions in 240,000 games by 58,000 players in tournaments rated by the World Chess Federation (FIDE).
- This excluded the first 8 moves in any game—"book" openings.

Chess Ratings

Idea: The *points expectation* E for player P versus opponent(s) O should be a function of the difference(s) in ratings $\Delta = R_P - R_O$ alone.

$$\begin{array}{rcll} \Delta = 0 & \Longrightarrow & E = 50\% & \\ \Delta = 200 & \equiv & E \approx 75\% & \text{(one st.dev.)} \\ \Delta \to +\infty & \Longrightarrow & E \to 100\%. & \end{array}$$

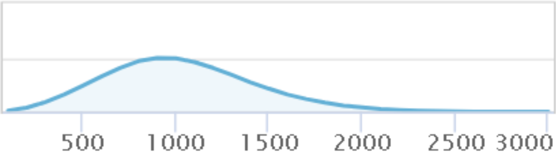
Sigmoid curve, such as USCF logistic curve:

$$E = \frac{1}{1 + \exp(-400\Delta \ln 10)}.$$

If your actual score exceeds (falls short of) your expectation then your rating goes up (down).

Elo Rating Examples

- Bobby Fischer hit **2800** on the US Chess Federation's Elo tabulation, **2785** on the FIDE list in July 1972.
- Current world champion Magnus Carlsen broke Garry Kasparov's record of 2851, reached peak of 2882. **Computers 3300+**.
- Current world #42 has 2703, world #100 has 2652.
- Formal "Master" designation in US 2200; "FIDE Master" more typical of 2300. Likewise "International Master" ≈ 2400, *Grandmaster* ≈ 2500, "strong GM" ≈ 2600.
- USCF uses 2000–2199 = "Expert," 1800–1999 = "Class A," 1600–1799 = "Class B" and so on.
- Distribution of online players on Chess.com—skewed low:



Intrinsic Chess Ratings (IPRs)

- Based on quality of your moves not results of games.
- Judged by chess programs stronger than all human players.
- Programs give *values* v in units of *centipawns* (cp).
- "Chatur Anga" (Four Strains of the army):
  - Pawn (peon), 100cp
  - Knight, Bishop: 300–350cp
  - Rook (boat): 500cp
  - Queen (vizier): 900–1,000cp.
  - Plus many other numerical measures of position structure...
- One virtue: many more data points of *moves* rather than results of *games*.
- (Will discuss IPRs later; focus on values now.)

The Value-Expectation Relation

$$E = \frac{1}{1 + \exp(-Bv)}.$$

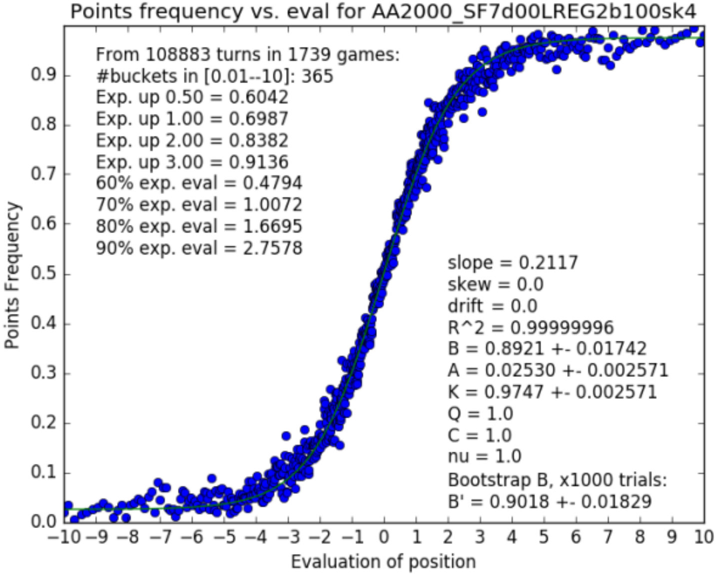
$$\boxed{\begin{array}{rcl} v = 0 & \implies & E = 50\% \\ B, v = 1 & \implies & E = \dfrac{1}{1 + 1/e} = \dfrac{1}{1.368\ldots} \approx 73\% \\ v \to +\infty & \implies & E \to 100\%. \end{array}}$$

Logistic curve, $B = B_R$ depends on the rating R.

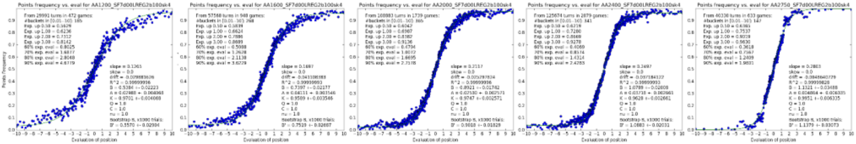
Refined to include small probability A of blundering away a "completely winning" game, giving a "generalized logistic" (Richards) curve:

$$E = A + \frac{1 - 2A}{1 + \exp(-Bv)}.$$

Example For Elo 2000 Rating



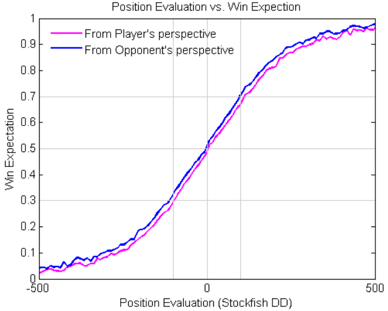
Slope as Rating Changes



- The slope B_R varies (linearly) with rating R.
- Hence mapping from v to E depends on R ("sliding scale").
- Google DeepMind's **AlphaZero** program uses only E in its move deliberations.
- In training by self-play it avoided the sliding-scale issue by "bootstrapping" its own B as it improved.
- But I have to model human players of all levels R in my tests.

We Can Already Make Some Inferences...

- The *same* factor B mediates both the chess program's value scale and the relation to rating.
- Suggests that *skill at chess is primarily the scale and vividness of one's perception of (differences in) value.*
- The frequency A of game-blowing blunders also varies with R.
- Given the position has value v, *ceteris paribus*, is it better if it is your turn to move or the opponent's turn? A "Murphy's Law":



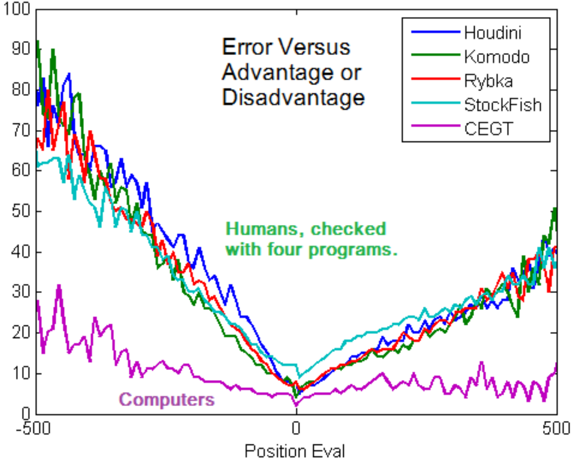
Law of Mass Sensitivity to Difference in Value

Conditioned on one of the top two moves being played, if their values in pawn units differ by...:

1. **0.01**, the higher move is played 53–55% of the time.
2. **0.02**, the higher move is played 58–59% of the time.
3. **0.03**, the higher move is played 60–61% of the time.
4. **0.00**, the higher move is played **57-59%** of the time.

- Last is not a typo. J.R. Capablanca and A. Alekhine had over 1,000 tied-top cases in their 1927 championship match.
- Almost 60% of the time, they played the move that Stockfish would list *first*—90 years later. ESP? Precognition?
- Similar 58%-42% split seen for any pair of tied moves. What can explain it?
- Will leave explanation as a "teaser" until the end...

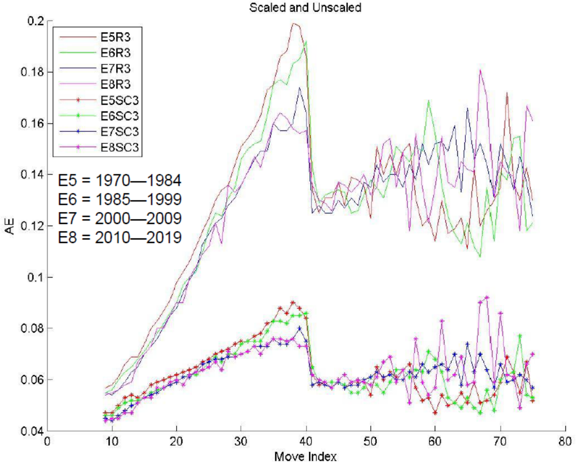
Law of Relative Perceived Differences in Value



Values can be scaled to flatten this out and conform more to E scale.

"Law" of Human Time Budgeting

Error By Move Number in Games



Effect of time pressure approaching Move 40 is clear.

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

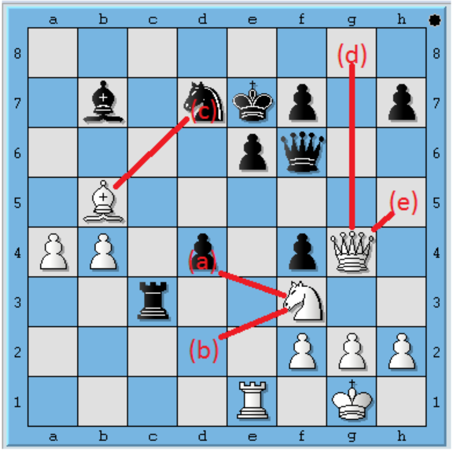
Chess and Tests

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

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

(source: itunes.apple.com)

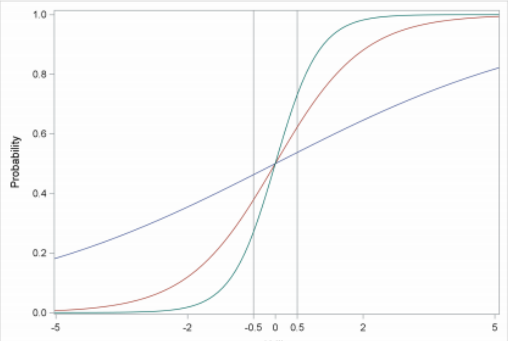
=



Item-Response Theory

- Students quantified by one aptitude parameter θ (“the” grade).
- Each test question q determines a curve $E_q(\theta) \equiv$ the likelihood of a person of skill θ getting it right.
- IRT posits this as always a Richards curve whose slope B is the sharpness of level that the question *discriminates*.

Figure 3 Item Characteristic Curves



Does Chess Conform to IRT?

- The analogue of getting a question right is playing exactly the move the computer judges best.
- Score = "Move-Match Percentage" (MMP or MM%).
- A second measure is how far off a person's wrong answers are.
- Or whether and how much partial credit is deserved for "close" answers.
- Use difference in value $v_1 - v_i$ to judge the ith-best move m_i.
- Scale down extreme differences (justified above) to define $\delta_i = \delta(v_1, v_i)$.
- Score = "Average Scaled Difference" (ASD).
- Also gives a *utility function* for possible moves.

Obstacles to Directly Testing IRT in Chess

- Would like to do a direct test of the same position π on players of many different rating levels R to see if the curve of the MM% frequency of "solving" π really is sigmoid.
- Many positions π occur in 1000s of games. . . but they are "book" - already known to most players. Like having the answers in advance.
- Chess.com keeps data on many puzzle positions. . . but it uses its own puzzle-rating system, not chess ratings, and it is even more heavily skewed to levels below 1100.
- So need to use *novel* positions—ones that are unique, never having occurred before. (My cheating tests use *only* these positions.)
- Can attempt to *cluster* positions π by similarity of δ_i mapping.
- Which "shape" produces the highest expectation of error (for any given R)? A kind of "Brachistichrone Problem" for chess.
- Otherwise, use my model's MM% and ASD projections directly.

The MM% Projection, 1600-to-2700 Levels

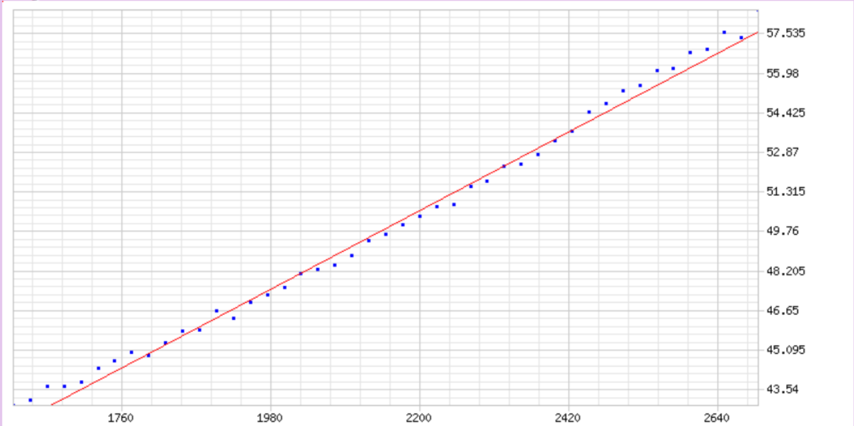
Function

$f(x) = 19.654619721630443 + 0.014057033867393376x$

R-Squared

$R^2 = 0.99303212012685$

Graph



Now Including 1025–1600, 2725–2800:

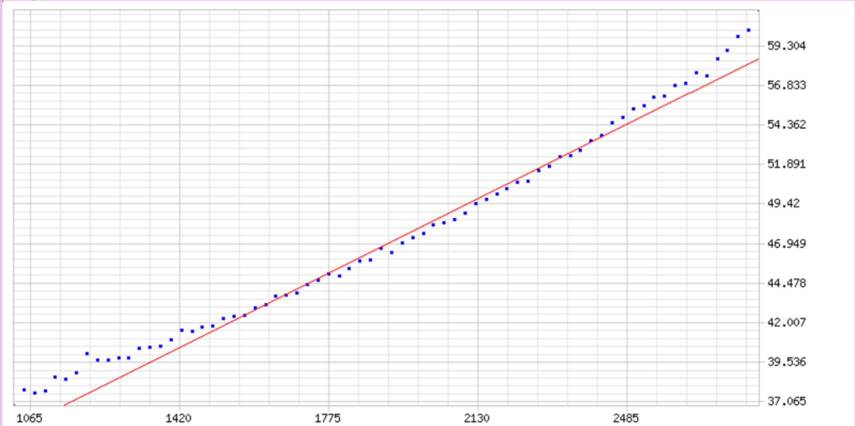
Function

$f(x) = 21.8651175524436\,6 + 0.013085915894893769x$

R-Squared

$R^2 = 0.97835646846452$

Graph



Quadratic Not Linear Law?

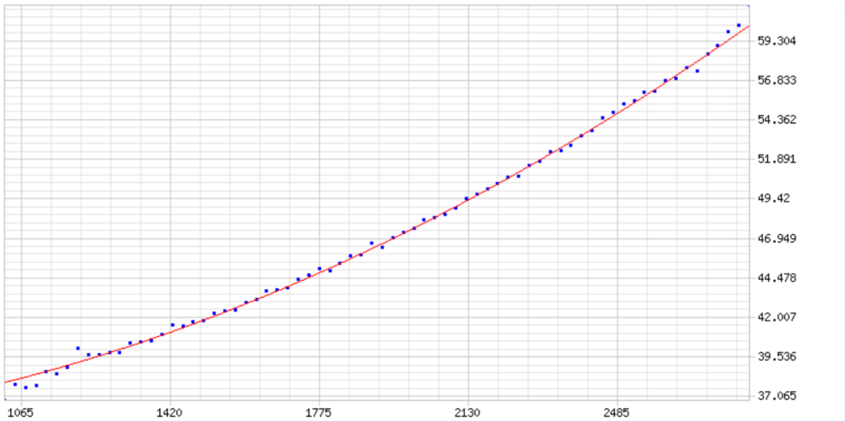
Function

$f(x) = 34.66026963709357 - 0.000243492414554713 68x + 0.0000033522002997568x^2$

R-Squared

$R^2 = 0.99779719205296$

Graph



Same With X,Y Axes Flipped...

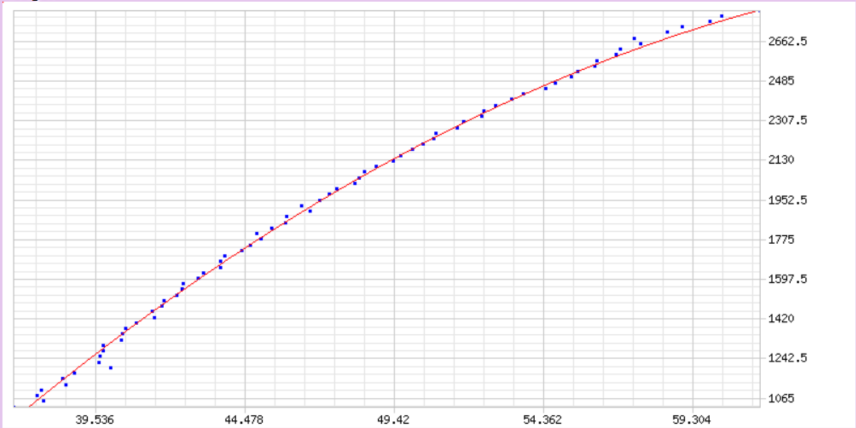
Function

$f(x) = -5224.3797654152 + 224.51739158320626x - 1.5285546730040955x^2$

R-Squared

$R^2 = 0.99814244490643$

Graph



...And Extended...

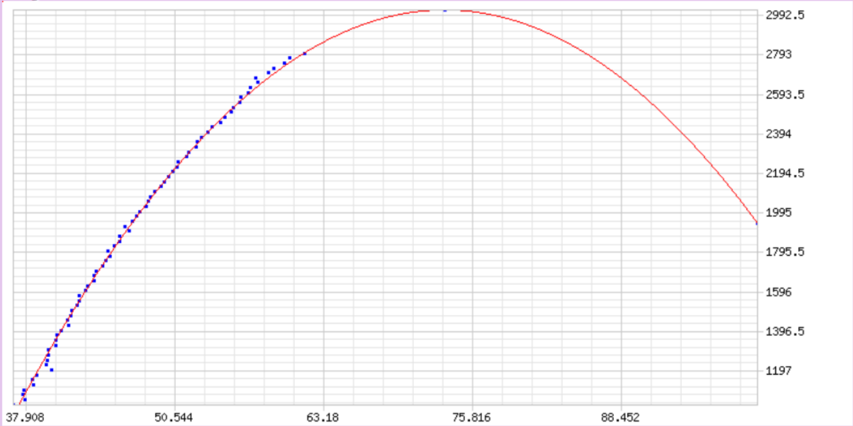
Function

$f(x) = -5224.3797654152 + 224.51739158320626x - 1.5285546730040955x^2$

R-Squared

$R^2 = 0.99825130391887$

Graph



Interpretations

- Seems ludicrous to think that 100% agreement with the chess program brings an amateur rating about 1950.
- Rather, an introspective conclusion: My methods and level of ("Single-PV") data-taking peter out toward Elo 3000.
- Computers match each other only 70–80% anyway.
- Most consider 3000 the watershed divide between the "human range" and the "computer range."
- My full model's "Multi-PV" data and equations seem to keep coherence up to about 3100.
- Can be so even if the level of Stockfish to *depth* at least 20 (up to 30 in positions with fewer pieces), i.e., searching 10 up to 15 moves ahead, is under Elo 3000.
- Analogy to catching particles with a river sieve.

Linear Law For ASD Looks Good...But...

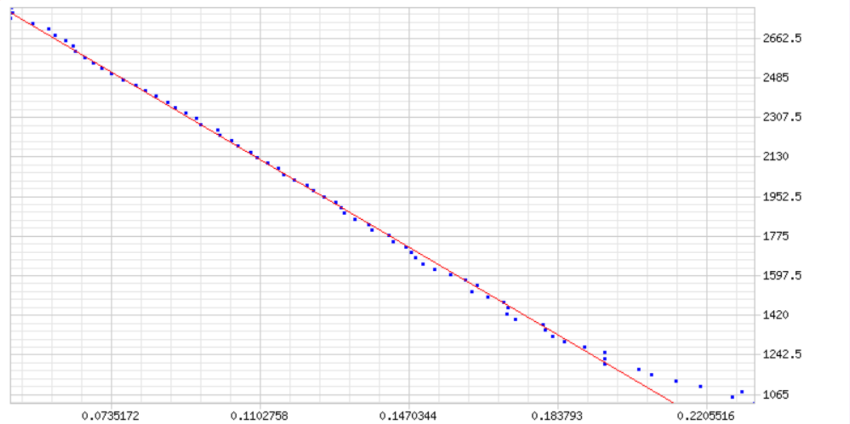
Function

$f(x) = 3298.02376454243 - 10688.627382908597x$

R-Squared

$R^2 = 0.99037759880581$

Graph



Quadratic Law Has Higher “Rating of Perfection”

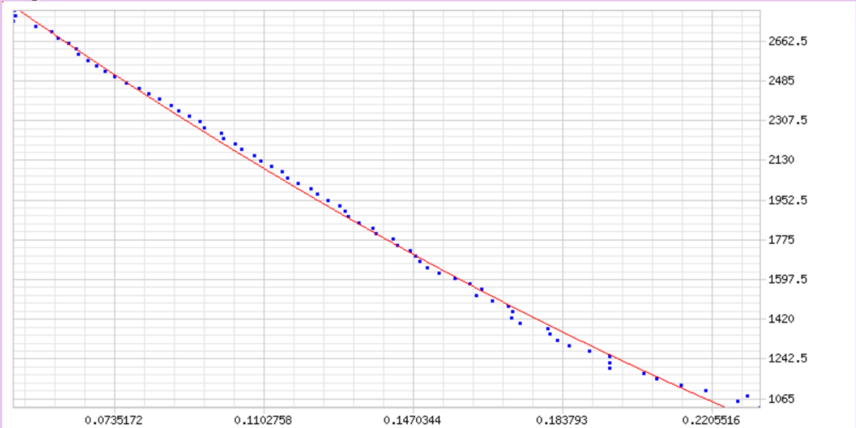
Function

$f(x) = 3462.663010383108 - 13884.604914850042x + 13415.403252920698x^2$

R-Squared

$R^2 = 0.99676481397797$

Graph



Multiplying By $4pq$ Recovers Good Linear Fit

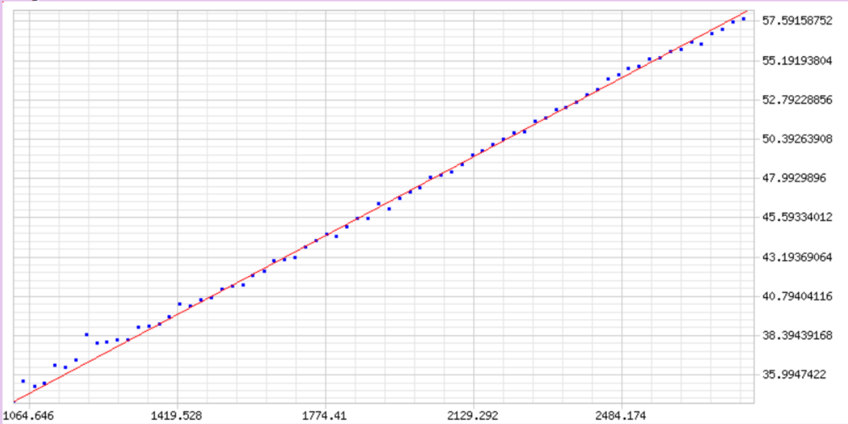
Function

$f(x) = 20.42277725109287 + 0.013578631028477313x$

R-Squared

$R^2 = 0.99732175601628$

Graph



Which Law Applies, and With What Horizon?

- The $4p^2q$ fit requires solving cubic equation to recover p.
- Equation becomes real-ly unsolvable when $p > 2/3$, so $4pq \approx 0.593$.
- Implies rating horizon of 2860, not 3000. Too low?
- Magnus Carlsen had 2860+ rating for 2-1/2 years but did not match 66.7%.
- So to re-pose the question: Is MM% quadratic?
- Any *non-linearity* can be a "game-changer" for scientific modeling, even if the local effects are small.
- Same questions for the law of ASD to skill.
- As currently constituted, my model's IPRs are primarily reflecting *accuracy*—avoidance of blunders.
- Can we reward *depth-of-thinking* directly?

Decision Model: Linear or Log-Linear or ...

- A "classical" *decision model* predicts the likelihood ℓ_i of a decision outcome m_i, which becomes its forecast probability p_i after normalization, in terms of its *utility* u_i to the decider.
- *Linear model* writes $\ell_i = \alpha + \beta u_i$.
- If utility is relative to optimum, so $u_1 = 0$, then $\ell_1 = \alpha$.
- *Log-linear model* (multinomial logit) puts $\log p_i = \alpha + \beta u_i$.
- Largely won 2000 Economics Nobel for Daniel McFadden.
- Then p_i is obtained by normalizing the likelihoods (e^α drops out)

$$L_i = \exp(\beta u_i), \quad \text{so} \quad p_i = \frac{\exp(\beta u_i)}{\sum_i \exp(\beta u_i)}.$$

- Has its own name: *Softmax*.
- So which law holds in chess: linear or log-linear?

Evidence for Neither: Needs "LogLogRadical" Model

Log-log-linear equation:

$$\log\log(1/p_i) - \log\log(1/p_1) = \beta u_i$$

yields

$$p_i = p_1^{L_i} = p_1^{e^{\beta u_i}}.$$

My deployed model inverts β as $1/s$ where s stands for *sensitivity*, and makes utility nonlinear with a second parameter c (for *consistency*):

$$p_i = p_1^{L_i} = p_1^{e^{\left(\frac{\delta(m_1,m_i)}{s}\right)^c}}.$$

Triple-decker exponentiation. *Is it a natural law?* Or an *unnatural* law?

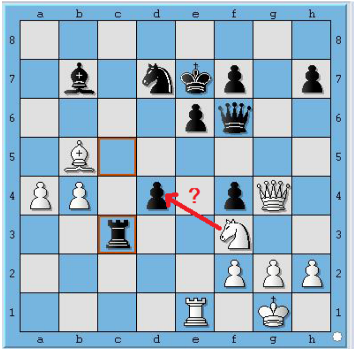
Check of Log-Linear Model: London 1883 Tmt

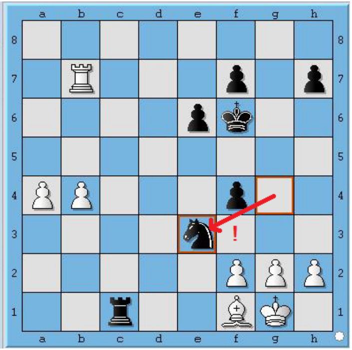
Rk	ProjVal	Actual	Proj%	Actual%	z-score
1	4870.99	4871.00	47.34%	47.34%	z = +0.00
2	1123.22	1729.00	10.94%	16.85%	z = +19.88
3	633.30	951.00	6.21%	9.32%	z = +13.27
4	459.83	593.00	4.56%	5.88%	z = +6.44
5	370.58	410.00	3.72%	4.11%	z = +2.11
6	311.98	295.00	3.16%	2.99%	z = -0.99
7	270.56	247.00	2.75%	2.51%	z = -1.46
8	239.36	197.00	2.44%	2.01%	z = -2.79
9	214.30	169.00	2.19%	1.73%	z = -3.15
10	193.93	104.00	1.99%	1.07%	z = -6.57

With LogLog-Radical Model (first line is MM%)

Rk	ProjVal	Sigma	Actual	Proj%	Actual%	z-score
1	4871.02	47.02	4871.00	47.34%	47.34%	z = -0.00
2	1786.89	37.32	1729.00	17.41%	16.85%	z = -1.55
3	929.87	28.60	951.00	9.11%	9.32%	z = +0.74
4	589.93	23.29	593.00	5.85%	5.88%	z = +0.13
5	419.35	19.84	410.00	4.21%	4.11%	z = -0.47
6	315.24	17.32	295.00	3.19%	2.99%	z = -1.17
7	246.68	15.39	247.00	2.51%	2.51%	z = +0.02
8	198.71	13.85	197.00	2.03%	2.01%	z = -0.12
9	161.54	12.52	169.00	1.65%	1.73%	z = +0.60
10	134.18	11.43	104.00	1.38%	1.07%	z = -2.64

The Deepest Mental Influence?





Values by depth of search:

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

Measuring "Swing" and "Heave"

- A move that initially looks best but whose value *swings down* on deeper reflection is a powerful *trap*.
- This one caught out Vladimir Kramnik in 2008 loss to Anand.
- Note also two moves are tied for equal-top value (**0.00** difference).
- The second-listed was more-often viewed as inferior.
- Computer chess programs use *stable* sorting—so it never becomes first unless viewed as strictly superior.
- Non-parapsychological explanation of 57–59% phenomenon.
- Dr. Biswas formulated a numerical measure ρ of the *swing* in value across depths—and showed far higher influence than I'd suspected.
- And that the depth of exposing mistakes grows linearly with skill rating R. Better players commit deeper errors.
- New model parameter h (for nautical "heave") multiplies ρ.

Interpretations and Modeling

- Operative Q on Depth of Thinking is not "what do you decide?" but

  "when and why do you decide to stop thinking?"

- So h could measure tendency to act prematurely.
- The "Perceived Utility" equation can be modeled like so:

$$u_i = -\frac{\delta(v_1, v_i) + h \cdot \rho(m_i)}{s},$$

  with either or both terms raised to the "radical" power c.
- This formulation makes h give the player's relative attention to the "subjective" value $\rho(m_i)$ versus the objective value v_i.
- So $h < 1$ means objective has higher influence, $h > 1$ subjective.
- Which one wins? We're human, right? Actually not clear...

Diverging Results and Difficulties of Control

- Fitting to equate actual and projected MM% and ASD typically yields $h > 1.5$.
- Whereas fitting by Maximum Likelihood Estimation (MLE) gives $h < 0.5$.
- Problem is MLE fitting gives diverging s, c values too and badly biases the MM% and ASD estimators.
- Equation fitting often gives *great* cross-check results... but also often fails to give a solution at all... or gives multiple solutions.
- Even when it works, the solutions destroy the previous uniform progression of s, c with rating R.
- The minimization landscape with just the s, c parameters is benign (a "canyon") but adding h creates "badlands" of non-local minima.
- Currently trying to have s, c touch components of ρ directly and add parameters that preserve the "canyon" shape.

Conclusions: Natural Laws and Mental Tuning

- Logistic-Curve Laws govern *expectation* from both *skill* and *value*.
- Relative Perception of Value—allows greater mistakes.
- Time Management Failings—complicate the modeling task too!
- MM% Agreement Law—linear or nonlinear?
- Value Swings and Decision Stopping Time—how best to model?
- *Predictive Analytics* is supposed to handle factors like these.
- But need to self-scrutinize one's modeling—to get it into tune.
- And need to be skeptical of the data used—to know the validity range of the data.
- Currently-deployed model has conservative fallback settings.
- Continued research and trials will hopefully give brighter light—and sharper guidance for our own mental fitness.