Reforming the FIDE Elo System: A Data-Driven Approach to Mitigating Rating Deflation in Chess

Yichen Han (International Master)
Magdalen College School, Oxford, United Kingdom

***Abstract*—The accurate assessment of player strength is essential for competitive chess, ensuring equitable competition, maintaining competitive integrity, and sustaining the sport's global ecosystem. The Fédération Internationale des Échecs (FIDE) continues to employ the Elo rating system—a probabilistic framework developed in the mid-20th century that, despite its historical significance and widespread adoption, has significant limitations in contemporary chess environments. Most critically, rating deflation has emerged as a fundamental challenge: player ratings systematically underestimate true playing strength, creating cascading effects throughout the competitive landscape. This deflation distorts tournament seeding, compromises predictive validity, and exacerbates inequalities between elite players and the broader professional community. This study presents a comprehensive quantitative analysis of over one million chess games to evaluate the empirical shortcomings of the current FIDE Elo system. Through rigorous statistical analysis spanning multiple temporal periods, we identify the structural factors driving persistent rating deflation and propose targeted modifications to the system's core mechanics. Our refined rating model incorporates adjustments to both the probability function and K-factor dynamics, designed to fix deflationary pressures while preserving essential qualities of fairness, stability, and predictive accuracy. These findings provide a robust, data-driven foundation for modernizing chess rating methodologies, with significant implications for player evaluation, tournament organization, and the sustainability of the global chess ecosystem. Code and data are available at https://github.com/IMYichenHan/FIDE-Analysis.**

I. Introduction

The measurement of competitive strength in chess serves as the cornerstone of tournament organization, competitive fairness, and the sport's institutional legitimacy worldwide. Since the adoption of the Elo rating system by the FIDE in 1970, this probabilistic framework has quantified player strength, determined tournament pairings, and governed access to elite competitions. Originally designed by physicist Arpad Elo, the system employs mathematical models that predict game outcomes based on rating differences and subsequently adjust player ratings according to actual results [1]. While the Elo system represented a revolutionary advance over its predecessors, its mathematical foundations have remained largely static for over five decades—a period that has witnessed major changes in chess training methodologies, competitive structures, and the global player population.

Recent data shows critical weaknesses in the FIDE Elo system, with rating deflation emerging as the most pressing concern. This phenomenon manifests when player ratings systematically fail to reflect actual playing strength, creating a persistent downward drift across the rating spectrum. The consequences extend far beyond statistical accuracy: deflation distorts tournament structures, skews prize fund distributions, and disproportionately impacts professionals outside the elite tier. While top players can insulate themselves through participation in closed round-robin events with similarly rated opponents, lower-tier professionals face a fundamentally different reality. They must compete in open tournaments against ambitious, systematically underrated opponents—resulting in rating losses even when performing at or above their nominal strength level [2].

The severity and urgency of rating deflation become apparent through analysis of FIDE ratings over time. Over the past fifteen years, we observe a pronounced decline in the number of players exceeding critical rating thresholds (2700 and 2650), despite consistent growth in global tournament participation. The impact on specific demographics proves particularly concerning: the number of female players rated above International Master level (2400+) and junior players exceeding Grandmaster level (2500+) has decreased significantly, especially following the COVID-19 pandemic. Current figures for these cohorts represent their lowest levels in over a decade—a troubling trend given substantial global investments in youth and women's chess development. These patterns indicate that the decline stems not from talent scarcity but from systemic flaws within the rating methodology itself.

Furthermore, game-level analysis reveals growing divergence between the Elo system's predicted outcomes and observed results. Grandmaster-level players (2500-2599) now achieve significantly lower win rates against FIDE Master-level opponents (2300-2399) compared to a decade ago, with success rates falling from 62% in 2014-2015 to 51% in 2024-2025. This erosion reflects fundamental shifts in the chess ecosystem: widespread access to advanced training tools, sophisticated engines, and large opening databases has elevated the competitive capabilities of lower-rated players beyond what legacy rating models anticipate.

This study addresses these challenges through a detailed analysis of the FIDE Elo rating system. By examining over one million recent game outcomes, we identify the structural factors contributing to persistent rating deflation and propose targeted modifications to the system's core mechanics. Our objective is to develop a refined rating framework that corrects systemic biases while preserving the essential qualities of fairness, predictive accuracy, and long-term stability. Through careful re-examination of the mathematical foundations underlying chess ratings, this work aims to support the evolution of a system that remains fundamental to global chess governance and competitive integrity. We accomplish this by proposing

three key modifications:

1) **Updating the probability function**: By recalibrating the outdated expected score formula to reflect contemporary competitive dynamics more accurately, we aim to mitigate rating deflation.
2) **Dynamic K-Factor formula**: By implementing a continuous linear function for the K-Factor, we account for the heterogeneity of rating dynamics across different rating bands while eliminating opportunities for rating manipulation and addressing problematic edge cases.
3) **Increased frequency of rating updates**: By transitioning from monthly to bi-monthly rating updates, we eliminate rating stacking phenomena and ensure that published ratings more accurately reflect players' current strength.

II. Related Work

Rating systems fulfill the essential function of quantifying relative skill through standardized numerical values, enabling meaningful competition across diverse contexts, geographic regions, and time periods [3]. While chess represents the most prominent application of such systems, similar frameworks have been successfully adapted for various competitive domains, from traditional sports like tennis to modern digital environments, including online gaming and matchmaking platforms. The evolution of chess rating systems provides valuable insights into both the challenges of accurate skill assessment and the continued efforts to improve.

A. Harkness System

The foundation of modern chess ratings began with Kenneth Harkness, a Scottish chess organizer who developed the first large-scale rating system formally introduced in Chess Review (Harkness, 1942). The United States Chess Federation (USCF) adopted the Harkness system from 1950 to 1960, implementing a straightforward percentage-based calculation methodology (Harkness, 1942). Under this system, tournament participants' ratings were adjusted based on their performance relative to the average rating of their opponents (Ravg). Players achieving exactly 50% scored received the average competition rating as their new rating, while those exceeding or falling short of 50% gained or lost 10 points per percentage point of deviation:

$$R_{\text{new}} = R_{\text{avg}} + 10 \times (S - 50), \tag{1}$$

where S represents the player's score percentage. An example calculation is provided in Section V.

Despite its innovative approach for the era, the Harkness system contained fundamental flaws that limited its long-term viability. Most critically, the system completely disregarded a player's existing rating when calculating adjustments, creating extreme vulnerability to volatility from anomalous tournament results [1]. Neither established skill level nor historical consistency influenced rating updates—meaning even accomplished Grandmasters performing poorly in a single event could see their ratings plummet to levels suggesting far weaker playing strength. This inherent instability rendered the system inadequate for cumulative, long-term player assessment.

B. Elo System

The revolutionary advancement in rating methodology arrived in 1960 when Arpad Elo, a Hungarian-American physics professor and chess master, developed the framework that would become the global standard. Building upon the foundation laid by Harkness, Elo introduced two key improvements. First, the implementation of logistic probability curves to calculate expected outcomes based on rating differences. Secondly, the adoption of progressive rating updates after individual games rather than entire tournaments [1]. These modifications created a more responsive and mathematically sophisticated system that better reflected the dynamic nature of competitive play.

The USCF promptly adopted Elo's system in 1960, replacing the Harkness framework. A decade later, FIDE followed suit in 1970, establishing Elo's methodology as the official international standard. The impact of this adoption extended far beyond chess, as the Elo system's elegant mathematical framework has since influenced rating models across numerous competitive fields, from professional sports to esports. The system's core mechanics, including its probability formula and update mechanisms, continue to form the backbone of modern chess ratings (detailed explanation provided in Section III).

C. Chessmetrics

Recent technological advances have enabled the development of alternative rating approaches that leverage computational power unavailable to earlier systems. Chessmetrics, developed by Jeff Sonas [4], represents a significant departure from traditional outcome-based ratings. Unlike the Elo system's exclusive reliance on game results, Chessmetrics employs sophisticated computer engine analysis to evaluate move-by-move quality, enabling unprecedented capabilities for cross-era player comparisons.

This computational approach addresses historical gaps in rating coverage. For instance, José Raúl Capablanca, the third world chess champion who reigned from 1921 to 1927, never received an official FIDE rating due to the Elo system's establishment decades after his career. Traditional methods cannot meaningfully compare Capablanca's strength to modern players like Magnus Carlsen. However, through detailed analysis of preserved games, Sonas estimated Capablanca's peak rating during his reign at 2813—providing context against Carlsen's peak of 2889 achieved in 2014.

While Chessmetrics offers powerful capabilities for historical analysis and statistical research, practical limitations prevent its adoption for large-scale rating applications. The computational resources required to analyze every game played by millions of active chess players worldwide remain prohibitive. Additionally, most amateur players lack sufficient recorded games to generate statistically reliable estimates. Thus, despite its analytical value, Chessmetrics remains unsuitable for widespread implementation in active competitive play.

Despite these innovations and ongoing discussions within FIDE, no alternative system has successfully challenged Elo's position as the universal standard for over-the-board chess. This persistence reflects both institutional inertia and the genuine difficulty of designing a model that optimally balances accuracy, fairness, simplicity, and computational feasibility for global implementation.

III. The FIDE system: A Closer Examination

The FIDE rating system represents a sophisticated implementation of Arpad Elo's probabilistic framework, employing rigorous statistical methods to quantify player strength across the global chess community [1]. Understanding its mechanics is essential for appreciating both its historical success and current limitations.

The system establishes a minimum base rating of 1400, with no formal upper limit—though the highest rating ever achieved stands at 2882, accomplished by Magnus Carlsen in May 2014 (FIDE Chess rankings 01-05-2014). As of July 2025, Carlsen remains the highest-rated active player with a rating of 2839. Notably, more than eleven years after Carlsen's peak achievement, the chess world has moved progressively further from the 2900 threshold rather than approaching it—a trend that warrants careful examination.

At its foundation, the Elo system calculates expected game outcomes between players using probability tables derived from rating differences. These tables assign values between -1 and +1 corresponding to both the probability of each result and the actual game outcome. The system's elegant symmetry ensures that rating points gained by one player exactly equal those lost by their opponent, maintaining mathematical balance across the entire rating pool (exact formulas and derivations provided in Section IV).

When players of different strengths compete, the system adjusts for skill disparities by awarding fewer points to the higher-rated player for victories while imposing greater penalties for losses. For example, in a game between Player A (rated 2400) and Player B (rated 2000), the probability of Player A winning approximates 0.92 [1]. Consequently, Player A gains only +0.08 rating points for victory while Player B loses -0.08. Conversely, should Player B achieve the unlikely victory (probability 0.08), they gain +0.92 points while Player A loses -0.92.

This zero-sum design ensures conservation of rating points across the population, theoretically preventing systematic inflation or deflation. However, as our analysis will demonstrate, real-world implementation reveals significant deviations from this theoretical equilibrium.

A. K-Factor

The K-factor serves as a critical multiplier that determines the magnitude of rating changes after each game. While FIDE has historically employed various K-factor values (including 25, 15, and 30), the current implementation uses three distinct categories based on empirical research [1], [4].

This differentiated approach reflects the empirical reality that junior and amateur players exhibit greater performance volatility, while elite players demonstrate more consistent strength levels. The K-factor thus modulates the system's responsiveness to match expected player development patterns (detailed mathematical formulas provided in Section V).

K Factor	Requirements
40	Juniors (less than 18 years) rated $<$ 2300, accommodating rapid improvement
20	Established juniors (2300-2400) or any adult player
10	Elite competitors ($>$ 2400), ensuring rating stability

TABLE I
THE THREE CURRENT K-FACTOR BANDS

In summary, for any two players, the rating difference determines their expected outcome via a probability table, which then outputs the gain/loss in points. This value is multiplied by the players' K-factors and added to their previous ratings to produce updated ratings. FIDE publishes these updates once per month.

B. The Problem of Rating 'Stacking'

FIDE's monthly rating update cycle introduces an unintended vulnerability known as rating stacking, whereby players can exploit the temporal lag between performance and official rating adjustments. This phenomenon occurs when players compete in multiple tournaments within a single rating period, effectively compounding their gains before ratings reflect their improved strength.

Consider Player A, rated 2000 with K-factor 40, who performs at a 2200 level across three tournaments in one month, earning +100 points per event. Because FIDE updates ratings monthly, their official rating remains 2000 throughout all three tournaments. By month's end, their new rating becomes 2000 + (3 × 100) = 2300—a result that defies logical progression. Had they consistently performed at 2200 strength, their rating should converge toward 2200, not overshoot dramatically.

Compare this to Player B under identical performance conditions but playing tournaments across three consecutive months. They gain +100 in month one (reaching 2100), then face diminishing gains in subsequent months as their rising rating reduces the point differential. Their rating correctly stabilizes around 2200, accurately reflecting their true strength.

This paradox exposes a fundamental flaw in FIDE's update frequency, creating opportunities for strategic manipulation of the rating system. While FIDE has introduced formulas to reduce K-factors for players participating in numerous monthly tournaments (see detailed formula in mathematics section), this approach creates new problems: it discourages active tournament participation (which FIDE should encourage) while failing to eliminate manipulation possibilities entirely.

The optimal solution—implementing more frequent rating updates, such as weekly calculations—would eliminate these distortions. However, FIDE has thus far been unable or unwilling to allocate the necessary computational and administrative resources for such a change.

C. Importance of FIDE ratings

FIDE ratings serve dual purposes within the global chess ecosystem, functioning as both recreational benchmarks for amateur players and critical professional credentials for serious competitors. For casual players, ratings provide tangible metrics for tracking personal improvement and setting achievement goals. However, for professional and aspiring professional players, FIDE ratings assume far greater significance: they determine access to international titles, govern invitations to prestigious tournaments, and ultimately influence financial viability within the sport. The credibility and accuracy of the rating system therefore carry profound implications for career trajectories, professional recognition, and competitive fairness throughout the chess world [2].

The pursuit of rating points and international titles imposes substantial financial burdens on emerging professionals. International tournament participation frequently requires extensive travel, with total costs often exceeding $2,000 per event when accounting for entry fees, accommodation, and transportation. Geographic disparities in rating pools compound these challenges significantly. Research demonstrates that Indian Grandmasters consistently outperform European players of similar ratings, revealing structural biases in regional rating dynamics [6]. These disparities create systematic disadvantages for players from developing chess nations, whose compressed local rating distributions tend to undervalue actual playing strength [6]. Our data additionally reveals notable declines in tournament participation over the past two years, particularly among top Grandmasters—a trend largely attributable to these mounting financial and structural barriers.

FIDE's international title system serves as the primary framework for recognizing chess achievement, with all titles directly or indirectly linked to rating thresholds:

Title	Rating Requirement
Candidate Master (CM)	2200
Woman's Candidate Master (WCM)	2050
FIDE Master (FM)	2300
Woman's FIDE Master (WFM)	2150
International Master (IM)	2400
Woman's International Master (WIM)	2250
Grandmaster (GM)	2500
Woman's Grandmaster (WGM)	2350

TABLE II
RATING REQUIREMENTS FOR DIFFERENT CHESS TITLES

For CM, WCM, FM, and WFM titles, players need only achieve the rating threshold once. The more prestigious IM, GM, WIM, and WGM titles require additional demonstration of consistent strength through three performance "norms"—tournament results meeting specific criteria. Despite these supplementary requirements, rating remains the fundamental determinant of title eligibility. This dependence proves especially consequential for the Grandmaster title, historically awarded to fewer than 1,800 individuals worldwide (FIDE, 2025), underscoring both its rarity and prestige.

The challenges of rating deflation impact elite players most acutely. For the top 100 players globally, ratings represent not merely reputation but livelihood. As deflation artificially suppresses ratings without reflecting decreased playing strength, these professionals risk losing ranking positions and, consequently, access to invitation-only tournaments that offer prize funds sufficient to sustain professional careers. Within this echelon, FIDE ratings constitute the sole meaningful metric of competitive standing. Maintaining a robust, fair, and empirically grounded rating framework proves particularly critical for this segment of the chess community.

Without addressing rating deflation, the gap between the protected elite—who operate within relatively insulated rating environments—and the broader population of deflation-affected professionals will continue widening. This trend carries serious implications beyond individual fairness. Reduced diversity in top-level competition risks diminishing audience engagement and viewership, potentially triggering declines in sponsorship, media coverage, and other revenue streams. Left unchecked, these dynamics could precipitate broader economic contraction throughout the chess ecosystem, compounding existing challenges for emerging talents while threatening the sport's long-term vitality.

IV. PROBLEMS WITH THE CURRENT SYSTEM

The Elo rating system adopted by FIDE in the 1970s has served as the global standard for evaluating chess players' relative strength for over five decades. Despite undergoing minor adjustments throughout this period, the system's core mechanisms have remained fundamentally unchanged. However, recent developments—particularly those intensified by the COVID-19 pandemic—have exposed significant limitations in the system's ability to accurately reflect contemporary playing conditions [4].

FIDE's head statistician, Jeff Sonas, made a notable contribution to addressing these concerns in July 2023 by proposing comprehensive reforms aimed at recalibrating the rating framework [4]. One such reform has already been implemented: raising the rating floor from 1000 to 1400 in March 2024 to compress rating categories. However, our research demonstrates that rating deflation persists despite this adjustment, indicating that more fundamental structural changes to the Elo system are necessary.

Current problems started with the COVID-19 pandemic, during which over-the-board tournaments ceased for more than a year. While physical competitions halted, many players—particularly juniors—continued intensive study and training through online platforms. This created a significant disconnect: players' actual strength improved substantially without corresponding increases in their official FIDE ratings, which only reflect over-the-board results. When tournaments resumed in late 2021 and early 2022, these players returned to competitive play with ratings that no longer accurately represented their abilities.

This mismatch between ratings and actual strength triggered widespread effects on competitive chess. Players with artificially deflated ratings (particularly juniors rated 1400-1500) frequently faced similarly underrated opponents, creating competitions where neither player's rating accurately reflected their true strength. Since the Elo system calculates rating changes based on expected outcomes derived from rating differences,

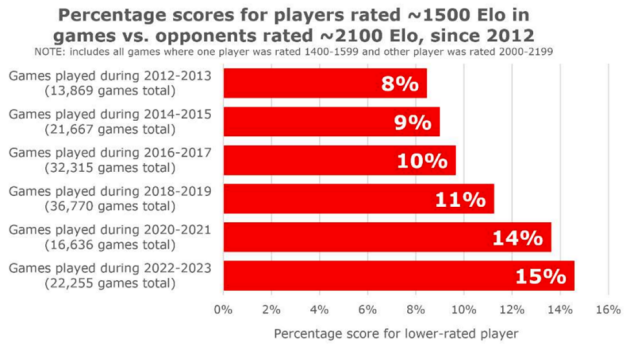

Fig. 1. The study Sonas conducted to show the change in results between 2100 and 1500 players

these mismatched games distorted rating point flows across the entire system. While some argue that such discrepancies should self-correct over time, this assumption only holds in closed systems with stable player populations [4]. The post-pandemic era has witnessed unprecedented growth in new players entering the rating pool, amplifying volatility rather than allowing natural correction [4].

The pandemic-era chess boom, catalyzed by cultural phenomena like Netflix's "The Queen's Gambit" (which attracted over 62 million households), further complicated these dynamics. The resulting surge in chess interest brought many new players into the FIDE system, typically entering at lower rating thresholds (1400-1600). As these newcomers entered competitive play, they exerted additional downward pressure on opponents' ratings, reinforcing deflationary cycles.

Theoretically, rating inflation from aging or inactive players retaining artificially high ratings should counterbalance deflation. However, practical evidence shows deflation accelerating more rapidly than any inflationary forces. This asymmetry occurs because skill acquisition among active players—particularly those with access to modern training resources—far outpaces skill degradation among inactive players. The pandemic's unique conditions, providing unprecedented time and motivation for intensive training, dramatically accelerated this trend.

The major impact of technology and artificial intelligence on chess training has fundamentally altered the competitive landscape. While top Grandmasters' peak strength has likely remained relatively stable over the past two decades, the average club player has undergone dramatic improvement. Modern training tools—including powerful local engines, NNUE (Efficiently Updatable Neural Network) engines, and large opening databases—have elevated typical club-level play far beyond historical norms. This creates a structural problem within the Elo framework: while the probability of a Grandmaster defeating a club player remains unchanged in the model's calculations, the actual skill gap has narrowed considerably. Consequently, Grandmasters lose rating points on average when facing lower-rated players who improve faster and adapt to new technologies more effectively.

This technological democratization has created structural inconsistencies in professional chess. Some established professionals can afford to participate in expensive closed tournaments where they avoid lower-rated opponents and maintain rating stability. Others, facing financial constraints, must compete in large open tournaments where they regularly face underrated club players, resulting in systematic rating losses despite maintaining or improving their actual strength.

Both our research and Sonas's investigations underscore the severity of these issues. Sonas's data reveals that lower-rated players now consistently exceed their expected performance as predicted by official Elo probability tables [4]. In 2012-2013, players rated 1500 scored just 8% against 2100-rated opponents—closely matching Elo predictions (FIDE, 2014). By 2022, that figure had nearly doubled to 15%. Fig. 1 shows the specific percentages each year.

Additional evidence emerges from analyzing the number of Grandmasters rated above the 2700 and 2650 thresholds since 2010, which both show a pronounced decline—a deeply concerning development for the sport's competitive ecosystem. These converging trends provide compelling evidence that the foundational assumptions underlying the Elo model no longer align with contemporary player demographics, development patterns, and the evolving dynamics of modern chess competition.

V. The Mathematics of Rating

This section contains further detail about the mathematical formulas mentioned in previous sections.

*a) **Harkness System**:* The Harkness rating system employs a straightforward calculation updating player ratings based on performance relative to 50% against average opposition:

$$R_{\text{new}} = R_{\text{avg}} + 10 \times (S - 50), \tag{2}$$

where S is the player's score percentage (%), and R_{avg} is the average rating of their opponents.

For example, if a player rated 2000 plays in an 11-round tournament against opponents averaging 2250 and scores 4/11 (36.36%), they are 13.6% below 50%, so their new rating would be: $2250 - 10 \times (50 - 36.36) = 2250 - 136 \approx 2114$.

*b) **Elo System**:* The Elo system (used by FIDE) models the expected score between two players based on their rating difference (Fig. 2), using a logistic curve.

For Player A (rating **R_A**) and Player B (rating **R_B**), the expected score for Player A is:

$$E_A = 1/(1 + 10^{(R_B - R_A)/400}) \tag{3}$$

Example:

If $R_A = 2200$ and $R_B = 2000$, then, $E_A = 1/(1 + 10^{-1/2}) \approx 0.76$. This means Player A is expected to score 76% against Player B.

*c) **Linear Elo Update (Simplified Form)**:* Arpad Elo also proposed a linear approximation in Fig. 3. The new rating can be calculated by the following equation:

$$R_{new} = R_{old} + K((W - L)/2) - (K/(4C)) \sum_i (D_i), \tag{4}$$

where R_{new} and R_{old} are the player's new and old ratings, respectively, D_i is the opponent's rating minus the player's

8. **The working of the FIDE Rating System**

The FIDE Rating system is a numerical system in which fractional scores are converted to rating differences and vice versa. Its function is to produce scientific measurement information of the best statistical quality.

8.1 The rating scale is an arbitrary one with a class interval set at 200 points. The tables that follow show the conversion of fractional score 'p' into rating difference 'dp'. For a zero or 1.0 score dp is necessarily indeterminate but is shown notionally as 800. The second table shows conversion of difference in rating 'D' into scoring probability 'PD' for the higher 'H' and the lower 'L' rated player respectively. Thus the two tables are effectively mirror-images.

8.1a The table of conversion from fractional score, p, into rating differences, dp

p	dp	p	dp	p	dp	p	dp	p	dp	p	dp
1.0	800	.83	273	.66	117	.49	-7	.32	-133.15	-296	
.99	677	.82	262	.65	110	.48	-14	.31	-141.14	-309	
.98	589	.81	251	.64	102	.47	-21	.30	-149.13	-322	
.97	538	.80	240	.63	95	.46	-29	.29	-158.12	-336	
.96	501	.79	230	.62	87	.45	-36	.28	-166.11	-351	
.95	470	.78	220	.61	80	.44	-43	.27	-175.10	-366	
.94	444	.77	211	.60	72	.43	-50	.26	-184.09	-383	
.93	422	.76	202	.59	65	.42	-57	.25	-193.08	-401	
.92	401	.75	193	.58	57	.41	-65	.24	-202.07	-422	
.91	383	.74	184	.57	50	.40	-72	.23	-211.06	-444	
.90	366	.73	175	.56	43	.39	-80	.22	-220.05	-470	
.89	351	.72	166	.55	36	.38	-87	.21	-230.04	-501	
.88	336	.71	158	.54	29	.37	-95	.20	-240.03	-538	
.87	322	.70	149	.53	21	.36	-102	.19	-251.02	-589	
.86	309	.69	141	.52	14	.35	-110	.18	-262.01	-677	
.85	296	.68	133	.51	7	.34	-117	.17	-273.00	-800	
.84	284	.67	125	.50	0	.33	-125	.16	-284		

Fig. 2. The probability table of the current FIDE rating system

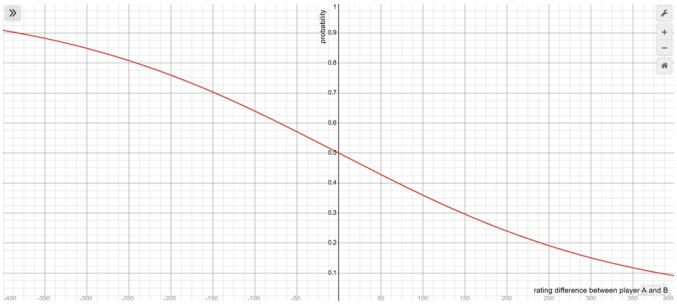

Fig. 3. The linear approximation graph of the expected result as the rating difference varies.

rating, W is the number of wins, L is the number of losses, $C = 200$ and $K = 32$. The term $(W - L)/2$ is the score above or below 0. $\sum D/4C$ is the expected score according to: 4C rating points equals 100%.

d) ***K-Factor Impact****:* The actual rating change per game is calculated by:

$$\Delta R = K \times (S - E). \tag{5}$$

Here, K is the development coefficient (often called the K-factor), S is the actual score (1 = win, 0.5 = draw, 0 = loss), E is the expected score. The expected score E can be derived by looking at the rating difference (D_i = opponent rating − player rating) and relating it to the expected result.

The K-factor controls how much the rating changes after each game:

- For juniors or new players, $K = 40$, making ratings more volatile (faster adjustments).
- For experienced players, $K = 20$.
- For elite players above 2400, $K = 10$, ensuring more stability.

For example, if a junior rated 2000 scores +2.0 points over several games, they would gain: $2.0 \times 40 = 80$ points, ending with a rating of **2080**. However, for elite players, the same +2.0 score would only yield: $2.0 \times 10 = 20$ points, due to their smaller K-factor.

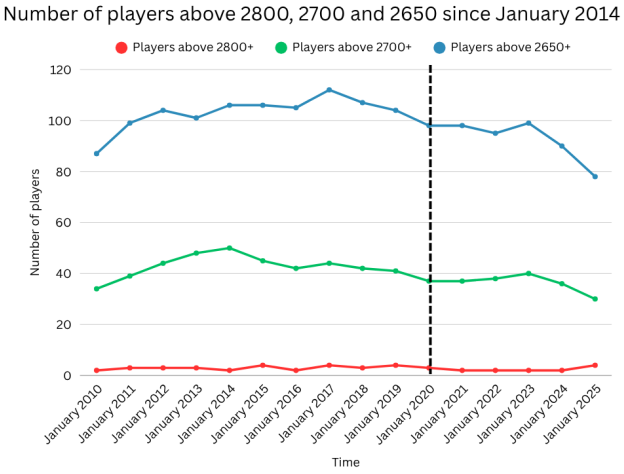

Fig. 4. A graph showing the number of players above 2800, 2700 and 2650 thresholds each January since 2010

VI. Problem Formulation

In this section, we have investigated 5 different categories to highlight the current problem with rating deflation.

A. Number of Professional players above the 2800,2700 and 2650 threshold

Fig. 4 presents compelling evidence of rating deflation at the highest levels of chess. The graph tracks players rated 2800+ (red line), 2700+ (green line), and 2650+ (blue line) each January since 2010. While the 2800+ category remains too small for meaningful statistical analysis (never exceeding four players), the 2700+ and 2650+ categories reveal clear patterns.

Both categories demonstrated steady growth from 2010 to 2017, after which numbers plateaued and entered gradual decline. The dotted line marks the COVID-19 pandemic's onset. While the pandemic years showed minimal change—largely due to suspended tournaments limiting rating movement—the post-pandemic period has witnessed a sharp decline. As of January 2025, both the 2700+ and 2650+ categories have reached their lowest levels in 15 years.

This trend unequivocally indicates top-level rating deflation. High-rated professionals increasingly lose rating points to lower-rated players, contracting the elite player pool. The implications prove particularly concerning given that tournament invitations typically depend on ratings. As the high-rated player pool shrinks, invitations concentrate among an ever-smaller group, enabling these players to protect their ratings through mutual competition. Meanwhile, players outside this elite circle, lacking access to closed tournaments, must compete in open events where they face lower-rated but often underrated opponents, accelerating their rating losses. This creates a self-reinforcing cycle: fewer top-rated players lead to fewer opportunities, which leads to further rating compression. Most alarmingly, the data suggests this decline is accelerating rather than stabilizing.

.

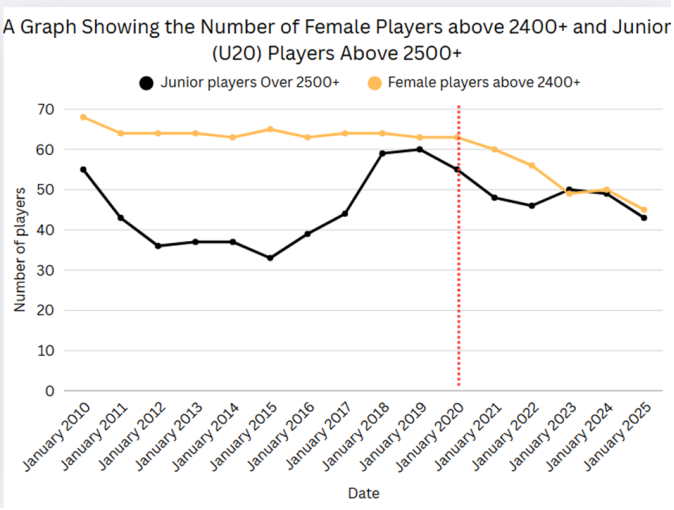

Fig. 5. A graph showing how the number of 2500+ junior players and 2400+ female players have changed since January 2010

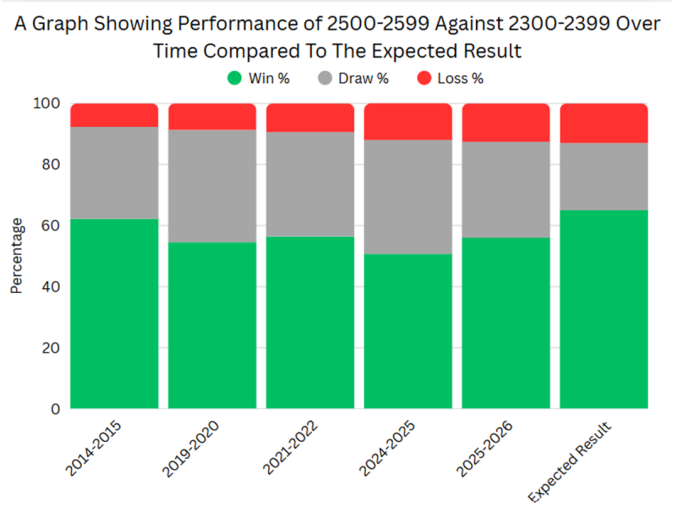

Fig. 6. The performance of 2500-2599 rated players against 2300-2399 in different time periods since 2014-2015, compared to the expected result

B. Number of Female Players above the 2400+ threshold and Junior Players above the 2500+ threshold

Fig. 5 reveals equally troubling trends among specific player demographics. The graph tracks female players rated above International Master level (2400+) and junior players (under 20) rated above Grandmaster level (2500+) since January 2010. Both categories showed consistent growth until 2020, when the COVID pandemic began. Since then, both have experienced marked declines. By 2025, the number of female players rated over 2400 has dropped by 33.8% compared to 2010 levels. Similarly, junior Grandmaster-level players have fallen to their lowest point in seven years.

These declines prove particularly troubling given the significant global investment in youth and women's chess development in recent years. Logic suggests these numbers should rise, reflecting successful development programs and increased participation. Instead, rating deflation has reversed expected progress, discouraging promising young and female players from pursuing professional careers as pathways to success become increasingly narrow. Without addressing deflation, continued investment in development programs will yield diminishing returns.

C. Performance of 2500-2599 Players Against 2300-2399 since 2014 compared to the Expected Result

Fig. 6 provides game-level evidence of rating deflation through performance analysis between Grandmaster-level players (2500-2599) and FIDE Master-level players (2300-2399) across different time periods. Based on 4,021 games over five years, the data reveals significant shifts in outcomes. Grandmasters' win rates have declined precipitously from 62% in 2014-2015 to just 51% in 2024-2025, while loss rates have risen from 8% to 13%.

These changes reflect the democratization of chess knowledge through technology. Strong amateur players now access sophisticated engines, large opening databases, and advanced training resources that enhance their resilience, reduce decisive mistakes, and increase drawing frequency. According to FIDE probability tables (FIDE, 2014), a 2500-rated player facing a 2300-rated opponent should achieve approximately 65% wins, 22% draws, and 13% losses to maintain rating equilibrium. While current loss rates match expectations, the dramatic increase in draws suggests amateur players, aided by engine preparation, have become significantly harder to defeat. Grandmasters, recognizing increased risk, may accept draws more readily.

Under current FIDE calculations, a 2500-2599 player facing 2300-2399 opponents would have gained 10 points per 100 games in 2014-2015. By 2024-2025, they would lose 65 points per 100 games—a 75-point swing. Even under expected conditions, they face a 45-point loss per 100 games. This reveals a fundamental flaw: the FIDE probability table, reasonably accurate in 2014, has become outdated and systematically unfair to higher-rated players.

This pattern extends across all rating bands. Mid-tier players face similar structural biases, with 2300-2399 players disadvantaged against 2100-2199 opponents. This "rating squeeze" effect drives systematic deflation as stronger players continuously lose points to slightly weaker opponents, compressing the entire upper rating distribution.

Individual Grandmasters have adapted by restricting participation to closed tournaments, avoiding exposure to dangerous lower-rated opponents. However, this strategy exacerbates inequality, creating two distinct professional chess worlds: a small protected elite maintaining rating stability, and the majority forced into open events with constant rating erosion. This division has undermined diversity and competitive fairness at the highest levels.

D. A comparison of average rating change per game at different rating differences

After collecting our data from August 2024 until August 2025, we calculated the average change of rating per game

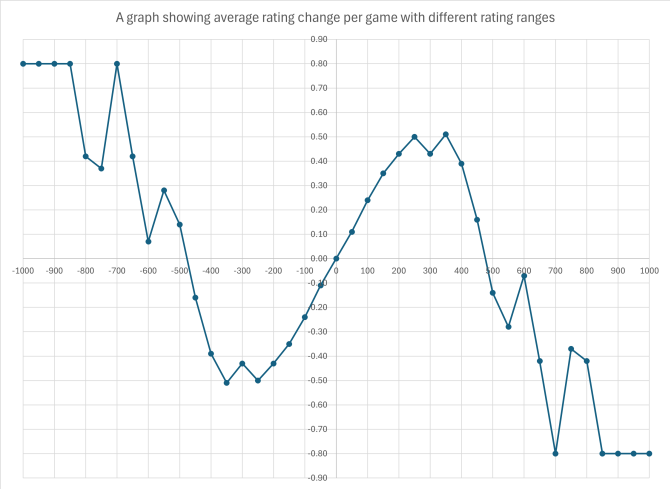

Fig. 7. A graph showing average rating change per game at different rating ranges

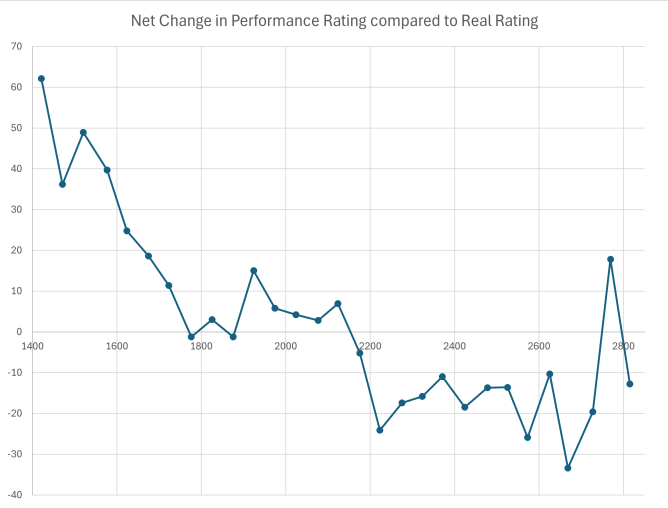

Fig. 8. Net Change in Performance Rating compared to Real Rating

at every rating difference from -1000 up to +1000. Fig. 7 highlights the average rating change per game at different rating ranges. The data illuminates fundamental flaws in the current Elo rating system. For example, a player rated 2500 playing someone 200 points higher (2700) will, on average, gain +0.5 points each game that they play. This finding is particularly surprising, as an ideal rating model should produce an average gain of exactly +0.0 per game, reflecting perfect calibration between expected and actual outcomes.

E. How different rated players perform compared to their actual rating

Additionally, we collected the performance rating from each rating 'band' from 1400 up to 2800. Fig. 8 shows the net change in Performance Rating compared to Real Rating. The x-axis shows the player ratings, while the y-axis shows the net difference between the players' rating and their performance rating in the past 12 months. From our data, we can highlight systematic rating deflation: players below 2200 are performing clearly above their level, whilst players above 2200 are consistently performing below their real rating, therefore consistently losing rating points. It is notable that there is actually an 'over performance' for players around the 2750-2800 rating range, which suggests that the 'elite' players, who participate in exclusive round-robin events, are isolated from the rest of the system and therefore unaffected by rating deflation.

VII. Solution Outline

This section outlines a comprehensive investigation designed to collect and analyze a representative dataset spanning the full range of competitive chess, from 1400 to 2899 ratings. The study assembled over one million games across carefully selected historical periods from chess-results.com, specifically: 2014-2015, 2019-2020 (pre-COVID), 2021-2022 (during COVID), 2024-2025 (post-COVID), and January-April 2025. These intervals enable both longitudinal analysis of rating behavior across an 11-year span and focused examination of COVID-19's impact on rating deflation.

A. Method

The investigation required developing a robust Python-based scraping tool capable of extracting and organizing over one million data points from FIDE-rated over-the-board classical games available on chess-results.com. The methodology categorized each game result (win, draw, or loss) into 50-point rating bands (e.g., 1400-1449, 1450-1499, up to 2850-2899). These aggregated results were then visualized and compared against Elo probability tables to identify systematic discrepancies. The source code for the Python scraper is provided on Github.

One notable limitation was that the chess-results database contains both missing tournaments and extraneous events outside the scope of the study (such as blitz tournaments). As a result, considerable manual effort was required to curate a representative set of tournaments for the scraper. While this introduces the possibility of selection bias and human error, efforts were made to apply systematic sampling across regions and formats, with the file of selected tournaments also available on Github, accessible via the bibliography.

After processing over one million games, the analysis successfully compiled 80 data tables documenting results by rating category and year. These datasets provide robust empirical support for the presence of sustained and accelerating rating deflation. The findings demonstrate that existing corrective mechanisms are insufficient to address this systemic issue, requiring immediate intervention to prevent further erosion of the professional player base.

B. Proposed Reforms

Based on extensive analysis, this study proposes two principal reforms to address rating deflation:

1) **Dynamic Probability Tables:** The current Elo probability table, unchanged for over five decades, fails to reflect modern competitive realities. The proliferation of chess engines and improved amateur preparation has fundamentally altered the competitive landscape. This

study proposes that FIDE implement a dynamic system whereby probability tables are recalibrated annually based on actual game results. Under this approach, FIDE would aggregate global game outcomes at year's end and generate new probability tables. Then, to account for differences within each rating band, they could also slightly shift K-Factors according to the volatility of players that year. Annual recalibration would guard against both deflation and inflation by continuously adjusting to actual performance patterns.

2) **Increased Rating List Frequency:** FIDE's monthly rating updates create opportunities for rating manipulation through "stacking," where players exploit static ratings across multiple tournaments. Most tournaments conclude within 3-13 days, making monthly updates unnecessarily infrequent. This study recommends implementing bi-monthly updates (1st and 16th of each month), substantially reducing opportunities for rating manipulation while maintaining administrative feasibility. This change would:
   - Reduce rating stacking opportunities by 50%
   - Provide more current ratings for tournament pairings
   - Better reflect the dynamic nature of competitive chess
   - Maintain administrative feasibility for FIDE

   This increased frequency represents a balanced approach—frequent enough to prevent major manipulation while avoiding the administrative burden of weekly updates. Combined with our other reforms, this change would significantly enhance rating accuracy and fairness.

Combined, these reforms would stabilize the rating system while preserving essential qualities of fairness and predictive validity. While implementation requires enhanced data processing capabilities from FIDE, the benefits extend throughout the chess ecosystem—from elite professionals to grassroots players. A revitalized rating system would enhance competitive integrity, strengthen the sport's appeal to sponsors and spectators, and support the development of future generations of chess talent.

VIII. Solution Implementation

Following thorough analysis of over one million FIDE-rated games spanning more than a decade, we have identified and developed solutions in order to have a more dynamic and accurate probability table.

A. The Probablity Table

Our comparison between current rating system expectations and actual game data reveals significant misalignment between FIDE's outdated probability formula and contemporary competitive reality. To address this disparity, we propose modifying the expected outcome formula to better reflect current playing conditions:

$$E_A = 1/(1 + 10^{(R_B - R_A)/475}). \tag{6}$$

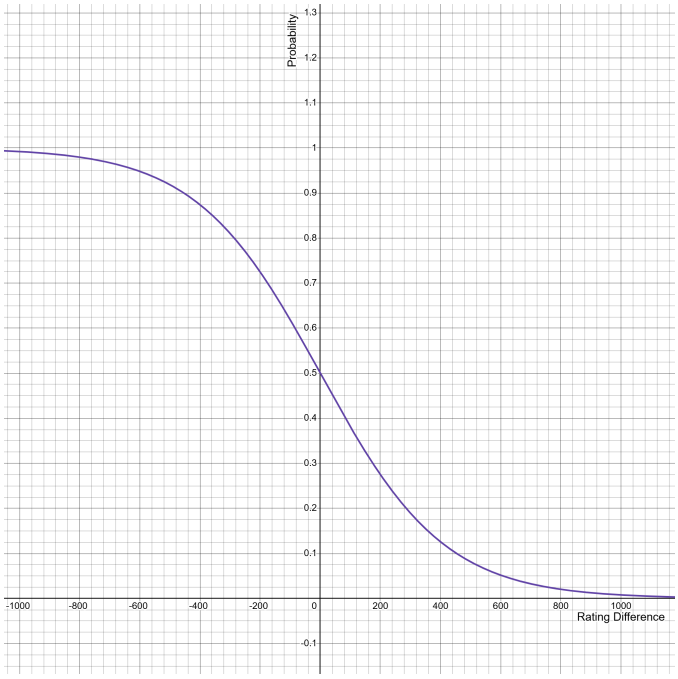

Fig. 9. A visualization of our proposed Expected Outcome

By increasing the spread parameter from Elo's original 400 to 475 in this logistic function, we maintain the mean at zero while adjusting the scale. This modification ensures higher-rated players receive slightly greater rewards for defeating weaker opponents, counteracting the deflationary pressure created by improved amateur play. Using the first-order Taylor approximation, this yields a linear relationship:

$$y = -0.001(R_B - R_A) + 0.5. \tag{7}$$

Under FIDE's current calculations, a 2500-rated player defeating a 2300-rated opponent gains +2.4 rating points with an expected outcome of 0.76. Our revised model increases this to +2.7 points, providing fairer compensation for the higher-rated player. This adjustment addresses the reality that technological advances have narrowed the practical skill gap between rating tiers, making lower-rated players more dangerous than historical models predict.

Our large probability table revision, which can also be found on Github, demonstrates how these adjustments better align with empirical game outcomes across all rating bands.

Fig. 9 shows a graphical representation of our new Expected Outcome formula. Fig. 10 shows a graphical representation of the linear approximation of our formula.

B. The K-Factor

The current FIDE system's rigid three-tier K-factor structure creates artificial discontinuities that can be easily manipulated. A player rated 2299 operates with K=40, while a player just one point higher at 2300 suddenly has K=20—halving their potential rating changes. This cliff-edge approach defies logic and invites strategic manipulation, with documented cases of players deliberately maintaining ratings like 2299

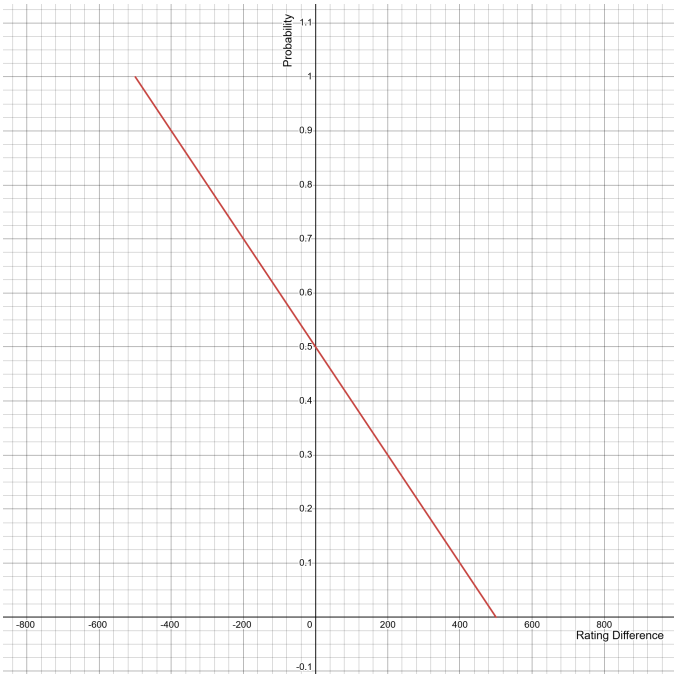

Fig. 10. Our first-order Taylor linear approximation

Fig. 11. our K-Factor (red) compared to the current K-Factor (blue) for junior players as rating changes.

to preserve higher K-factors before exploiting rating stacking opportunities.

We propose replacing this stepped system with a continuous linear function for players rated between 2000 and 2750:

$$K = (-1/25)R + 120, \tag{8}$$

Where R represents the player's current rating. This formula provides smooth transitions:

- Players below 2000: K = 40 (maintaining current junior/new player volatility)
- Player at 2400: K = 24 (compared to current K = 10)
- Player at 2500: K = 20 (compared to current K = 10)
- Players above 2750: K = 10 (preserving elite stability)

Junior players and those with fewer than 30 games retain K = 40 until reaching 2000, ensuring rapid initial adjustments. This continuous function eliminates manipulation opportunities while providing appropriate volatility across the rating spectrum. The slightly higher K-factors compared to the current system are justified by our more accurate probability tables—with better predictive models, larger adjustments help players reach their true ratings more efficiently. Fig. 11 shows a graphical representation of our new K-Factor formula.

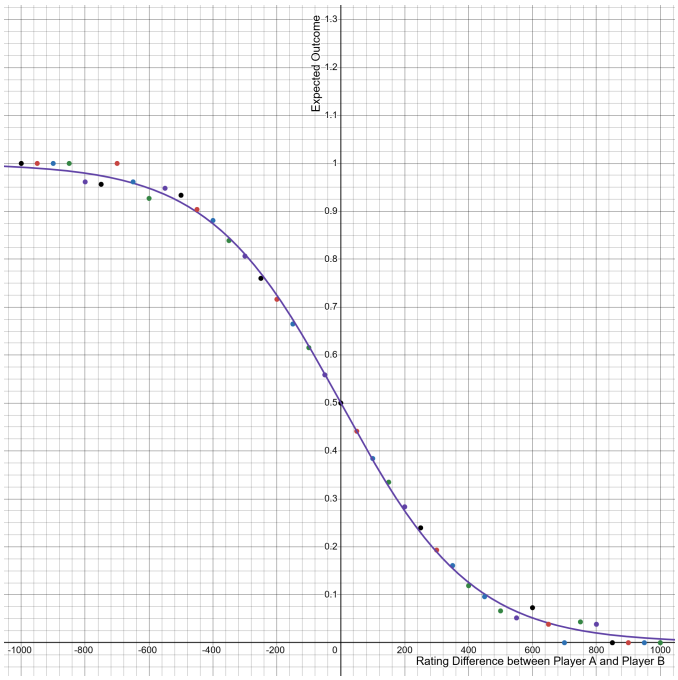

Fig. 12. A graph showing how our Expected Outcome formula compares to the data we sampled

IX. Analysis

While thorough testing of our reformed system requires real-world implementation, we conducted a statistical test where we tested our new Expected Outcome formula against the current FIDE system using a random sample of 116,906 games played between different rating bands from August 2024 to August 2025.

Our methodology calculated the deviation between each system's expected outcomes and actual game results across 50-point rating intervals. The analysis revealed that our modified probability function consistently provided better alignment with empirical outcomes, particularly in games between players separated by 200-400 rating points—precisely where current deflation proves most problematic.

Fig. 12 shows how our Expected Outcome formula compares to the data we collected. Fig. 13 shows how the current FIDE Expected Outcome formula compares to the data we

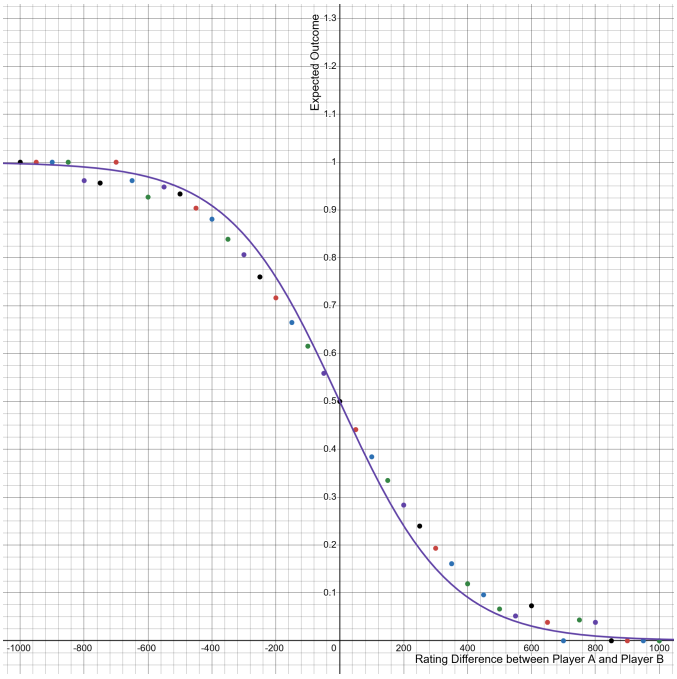

Fig. 13. A graph showing how the FIDE Expected Outcome formula compares to the data we sampled

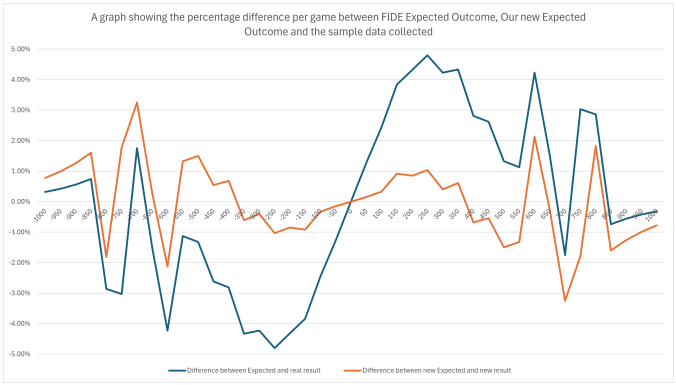

Fig. 14. A graph showing percentage difference per game between FIDE Expected Outcome, our new Expected Outcome and sample data collected

sampled. It is clear that our approximation demonstrates substantially closer alignment to the real data compared to FIDE's current Elo formula. Fig. 14 shows the percentage difference between our Expected Outcome and the actual sample data (orange line) and the difference between the current Elo FIDE Expected Outcome and the actual sample data (blue line). While most probabilities estimated by FIDE's formula deviate by 3% to 5% from the actual sample data, our formula has reduced the margin of error to 0% to 2%.

Additionally, while the FIDE Elo data has a standard sample deviation of 2.83%, ours achieves only 1.33%, demonstrating that our model is clearly more consistent with current sample results compared to the current FIDE Elo rating. The confidence level that our formula aligns more closely with the testing data than the Elo FIDE rating (smaller absolute error) is 99.999%.

X. Further Improvements

While our data-driven reforms would significantly improve the current rating system and reverse deflationary trends, we acknowledge that Elo systems exhibit complex long-term dynamics that cannot be fully predicted through statistical analysis alone. FIDE, with its superior resources and full data access, could refine our proposed formulas even further.

We recommend several avenues for continued enhancement:

Machine Learning Integration: As rating trends evolve annually, machine learning algorithms could continuously optimize probability tables and K-factors based on emerging patterns. This would create a truly adaptive system responding to changes in the chess ecosystem.

Regional Calibration: Our analysis revealed significant regional variations in rating accuracy. Implementing region-specific adjustments during the initial calculation phase could address these disparities while maintaining global compatibility.

Performance Volatility Modelling: Incorporating recent performance volatility into K-factor calculations could better distinguish between genuinely improving players and those experiencing temporary fluctuations.

Anti-Manipulation Measures: Beyond increased update frequency, algorithmic detection of suspicious rating patterns could flag potential manipulation for review.

Inactive Player Adjustments: Implementing rating decay for extended inactivity of, for example, over five years would prevent rating inflation from dormant high-rated accounts.

These enhancements would leverage modern computational capabilities to create the most accurate and manipulation-resistant rating system possible, ensuring chess ratings remain meaningful measures of playing strength in an evolving competitive landscape.

XI. Conclusion

This study analyzed structural flaws in the current FIDE Elo rating system, highlighting growing rating deflation—especially post-pandemic—through an empirical review of over one million games (2014–2025). Findings show the system underestimates elite strength and fails to reflect rapid club-level improvements driven by modern training tools. We proposed two core reforms: (1) annual recalibration of probability tables using real-world data, and (2) bi-monthly rating updates to curb manipulation and better track player development. These changes would restore fairness, accuracy, and trust in the system. Though implementation requires FIDE's commitment, the benefits—improved integrity, increased engagement, and stronger professional pathways—are significant. Ultimately, continuous, data-driven adaptation is essential to preserving chess's status as a global, competitive sport.

References

[1] Elo Arpad, "The rating of chess players, Past and Present" *https://gwern.net/doc/statistics/order/comparison/1978-elo-theratingofchessplayerspastandpresent.pdf*, 1978.

[2] Regan Kenneth and Haworth Guy, "Intrinsic Chess Ratings", *https://cse.buffalo.edu/ regan/papers/pdf/ReHa11c.pdf*, 2011.

[3] Glickman Mark E., “A comprehensive guide to chess ratings” *https://www.glicko.net/research/acjpaper.pdf*, 1995.
[4] Sonas Jeff, “Sonas Proposal: Repairing the FIDE standard Elo rating system”, *FIDE document*, 2023.
[5] Glickman Mark E., “An example of the Glicko-2 system”, *https://glicko.net/glicko/glicko2.pdf*, 2012.
[6] Berg Arthur, “Statistical Analysis of the Elo Rating System in Chess” *Chance, vol. 33, No. 3*, 2020.