Liverpool John Moores University

Analysis of racket-sport database rankings for behavioural effects and other anomalies

by

Arel Akaunu

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the
Astrophysics Research Institute

February 2026

Abstract

This dissertation evaluated whether behavioural factors produced systematic deviations between observed squash match outcomes and those predicted by the SquashLevels rating model, focusing on two candidate mechanisms: same-day fatigue and effort modulation in mismatches. An anonymised SquashLevels dataset (`sq_combined.csv`), precompiled via SQL joins across multiple relational tables, was analysed using deviation-from-expectation metrics derived from pre-match player levels and observed match dominance. For fatigue, players with at least two competitively weighted matches per day (`weighting` ≥ 0.75) were examined using a paired first-versus-last design ($n = 237{,}114$ player-day pairs), and a Wilcoxon signed-rank test assessed whether later matches showed systematic underperformance; the change was statistically significant ($p = 1.18 \times 10^{-37}$) but small and positive (median ΔDeviation $\approx +0.015$), providing no evidence of fatigue-related underperformance under this proxy. For effort modulation, matches were analysed from the stronger player's perspective and a log–log regression baseline related log(`RATIO`) to log(Level Ratio) ($n = 3{,}826{,}051$, $R^2 \approx 0.294$); residuals were compared across SquashLevels-aligned mismatch bands (1.0–1.5, 1.5–3.0, > 3.0), showing a negative median residual in extreme mismatches (approximately -0.050), consistent with reduced dominance relative to expectation. Overall, the findings partially supported the unifying thesis: fatigue effects were not practically evident within competitively weighted same-day sequences, whereas extreme mismatches exhibited systematic underperformance consistent with effort modulation, supporting the rationale for dampened rating sensitivity in highly unbalanced matches and indicating that behavioural adjustments were most defensible when targeted to contexts where the deviation signal was strongest.

Contents

List of Figures

Introduction

Context and Motivation

Ranking systems are essential in competitive sports for comparing players, determining seedings, and evaluating progress (Irons et al. 2014). Traditional ranking methods—such as tournament points or league tables—often fall short when players compete across loosely connected divisions or sporadic matchups (Csató 2021). To address these limitations, Elo-style systems have become widely adopted in sports such as chess, tennis, and online gaming, offering dynamic, match-by-match rating updates based on performance expectations and outcomes (Glickman 2011; Elo 1978).

In racket sports like squash, this need is especially pronounced. While the professional circuit uses point-based rankings, the vast majority of squash players—amateur, club, and league competitors—operate outside that structure. The SquashLevels system was developed to bridge this gap (*About the SquashLevels System*, no date). It ingests match data from league platforms, club systems, and manual entry by players to maintain a centralised, real-time rating system for all players regardless of region or level. This integration is illustrated in Figure 1.

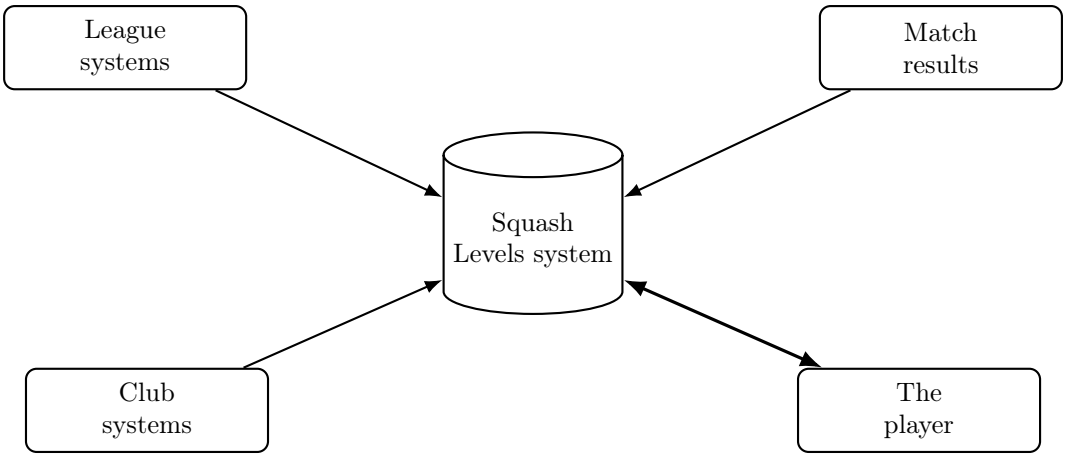

FIGURE 1: Core inputs to the SquashLevels system and the player interaction.

Unlike static point tables, SquashLevels continuously updates a player's Level after each match using an Elo-like algorithm. It factors in opponent strength, score margins, and the difference between expected and actual outcomes. The Level ratio between two players determines the predicted share of rallies won. Over five million results feed into this model, making it a powerful tool for evaluating and comparing squash performance in real time (*About the SquashLevels System*, no date).

Project Focus and Scope

While SquashLevels incorporates several refinements, such as discounting blowout matches and weighting recent results more heavily, its core algorithm still assumes that players perform at their true level in each match. In practice, however, this assumption does not always hold. Sports behaviour research has shown that physical and psychological factors can distort match outcomes, leading to rating anomalies not attributable to changes in skill.

This project examines two such behavioural anomalies:

- **Fatigue effects** — performance decline when players compete in multiple matches on the same day.
- **Effort modulation in mismatches** — situations where stronger players appear to reduce intensity in unbalanced contests.

Both cases challenge the assumption of constant effort and ability. Fatigue introduces temporary physical limitations, while effort modulation reflects strategic or social factors, such as conserving energy or avoiding embarrassment in lopsided matches. If unaccounted for, these behaviours could distort rating updates and undermine ranking fairness.

Relevant Literature

Past research has explored behavioural anomalies in various sporting contexts. In tournament formats, players have been observed to modulate effort to manipulate draws or conserve energy for later rounds (Krumer et al. 2023). Even in rating-based systems, psychological and strategic behaviour affects play. For example, studies in tennis and chess reveal that players adjust risk-taking based on ranking status, momentum, or opponent prestige (Holdaway and Vul 2021).

Fatigue effects are also well-documented in sports science. Repeated high-intensity matches can impair performance, especially in physically demanding sports like squash (Girard and Millet 2009). SquashLevels' documentation acknowledges this risk, noting that a player's rating may unfairly decrease after underperforming in a final match due to accumulated fatigue (*About the SquashLevels System*, no date). However, no formal adjustment currently exists for this.

Effort modulation is explicitly handled in SquashLevels for extreme mismatches. The algorithm applies an "effective effort" dampening factor when one player is more than three times stronger than the other, to avoid harshly penalising the stronger player for a close result. This built-in correction suggests awareness of effort variation.

Contribution and Report Outline

This dissertation contributes an empirical, data-driven analysis of fatigue and effort modulation in squash match data. Using the SquashLevels dataset, the project quantifies deviations between observed and expected outcomes under these two behavioural conditions. It tests whether the patterns match the rationale behind current SquashLevels adjustments and whether the deviations are statistically and practically significant.

Chapter 2 reviews the theoretical and algorithmic background of Elo-like systems and outlines the foundations of behavioural analysis in sport. Chapter 3 details the data sources and the methodological approach, including how deviations are defined and tested. Chapter 4 presents the results, and Chapter 5 interprets the findings in light of behavioural theory and the SquashLevels model. Chapter 6 concludes with limitations and implications for ranking fairness.

Background

Elo and Elo-like Systems

The Elo rating system, originally developed for chess, is a foundational method for modelling skill through pairwise outcomes. It updates a player's rating based on the surprise of the result: unexpected wins yield large gains, and expected wins yield small adjustments (Elo 1978). The update is based on the difference between the actual result and the expected probability of winning, computed from the rating difference between players.

This framework was extended through models such as Glicko, which introduced uncertainty tracking, and TrueSkill, a Bayesian system designed for multiplayer and team games (Glickman 2011; Herbrich et al. 2006). These systems allow for more flexible and responsive updates when player reliability is low or performances are erratic.

In sports applications, Elo-like systems have been enhanced to incorporate margin of victory, converting binary win/loss data into continuous metrics. This is especially relevant for squash, where scoreline dominance contains rich performance information. Margin-sensitive Elo variants have improved predictive power in sports like tennis and football (Kovalchik 2020).

SquashLevels Algorithm

SquashLevels is an Elo-inspired ranking system tailored for squash. Unlike basic Elo, it considers both the point and game scores to calculate a performance ratio for each match. It uses simulated match data to convert raw match results into a continuous performance scale, enabling fine-grained player comparisons even in non-professional contexts (*About the SquashLevels System*, no date).

The core model calculates an expected outcome from the level ratio:

$$\text{Expected Ratio} = \frac{\texttt{LEVELBEFORE}}{\texttt{OPPLEVELBEFORE}}$$

and compares this with the observed outcome (a simulated performance ratio based on match results). The difference between expected and actual determines the adjustment. This framework aligns well with our focus on behavioural anomalies, as we analyse deviation from expected results as a proxy for effort or fatigue-related effects.

Handling Behavioural Effects in SquashLevels

SquashLevels includes specific logic to handle behavioural variation. When the level ratio exceeds 1.5:1, the algorithm assumes that the stronger player may not exert full effort and applies an "effective effort" adjustment, reducing the expected dominance to avoid penalising the favourite for a closer-than-expected result. In extreme mismatches (>3:1), no adjustment is made if the stronger player wins as predicted (*About the SquashLevels System*, no date).

This effort-based dampening directly relates to the effort modulation hypothesis studied in this project. It acknowledges that in low-stakes, one-sided matches, stronger players may strategically reduce effort — a behaviour also observed in the literature on tournament manipulation and psychological adaptation in sport (Nieken and Stegh 2010).

For fatigue effects, the SquashLevels system does not currently adjust a player's rating based on match order or recovery time. By comparing performance deviations between first and last matches in a day, we can assess whether fatigue systematically affects outcome relative to model expectation.

These behavioural mechanisms (effort and fatigue) are not only theoretically grounded in the design of SquashLevels but also supported by findings in sport science, where effort pacing, cumulative fatigue, and motivation vary across match context.

Summary

SquashLevels provides a rich, dynamic model of squash performance that enables behavioural anomaly detection through its transparent, ratio-based updates. The system's architecture and design choices allow meaningful interpretation of deviations from expected results—making it an ideal foundation for this project's analysis of fatigue and effort-related behaviour in real match data.

Data and Methodology

Data Sources and Genesis

The primary data source for this project is the SquashLevels system – a real-time squash ranking platform that aggregates results from league systems, clubs, and player submissions (*About the SquashLevels System*, no date). The platform uses an Elo-inspired framework and stores match records in a relational database.

For this analysis, a consolidated dataset was created by the SquashLevels team prior to the start of the project. The dataset was constructed using SQL to join multiple core tables by a unique match identifier (`MATCHID`). Although this data preparation was not performed by the author, the final dataset includes all relevant player, match, and rating metrics in a cleaned, structured format suitable for analysis.

All data is fully anonymised, and no personal or identifying information about players was included in the dataset or its analysis. Each entry represents a single match between two players, with their performance and contextual information embedded.

Source Tables and Features

The dataset integrates fields from the following source tables:

- **sq_levels** (for both home and opponent players): `PLAYERID`, `MATCHID`, `LEVELBEFORE`, `LEVELAFTER`, `RATIO`, `CONFIDENCE`, and mirrored fields for the opponent: `OPPLEVELBEFORE`, `OPPLEVELAFTER`, `OPPRATIO`, `OPPCONFIDENCE`.
- **sq_matches**: `MATCHTYPEID`, `HOMETEAMID`, `AWAYTEAMID`, and `DATE/TIME` (derived from match datetime).
- **sq_strings**: parsed results including `GW` (games won), `GL` (games lost), `SCORE_TYPE`, `PTS_PER_MATCH`, and `FRAC_PTS_WON`.

- **sq_matchtypes**: `COUNTYID` and `WEIGHTING`, indicating match classification and importance.

The final dataset is structured with one row per match, including both players' ratings before and after the match, performance metrics, and match-level context. The `LEVELBEFORE` and `OPPLEVELBEFORE` fields serve as the basis for expected performance, while the `RATIO` and `OPPRATIO` represent actual match dominance, such that a ratio of 1.15 says that the player in this row played 15% better than their opponent.The `FRAC_PTS_WON` metric reflects point-based match dominance on a 0–1 scale. Match `WEIGHTING` (e.g., 0.5 for informal league matches, 1.0 for major tournaments) is used to filter for competitive play.

The dataset has undergone light cleaning and standardisation prior to this project, such as datetime formatting and parsing of result strings. No personally identifiable information was ever included in the analysis.

Methodological overview

Fatigue Effects

This analysis investigates whether players show signs of performance decline across multiple matches played on the same day, consistent with fatigue-related behavioural effects mentioned in the previous chapter. If present, such effects would represent as a systematic underperformance in later matches, relative to SquashLevels expectations according to levels of each of the players (see Figure 2 for details of the pipeline).

For each match, a performance deviation was computed by comparing the observed match outcome to the outcome predicted by SquashLevels ratings prior to the match. Specifically, the log-transformed difference was calculated as:

$$\text{Deviation} = \log(\texttt{RATIO}_{\text{observed}}) - \log\left(\frac{\texttt{LEVELBEFORE}}{\texttt{OPPLEVELBEFORE}}\right) \tag{1}$$

This deviation measures whether a player overperformed (> 0) or underperformed (< 0) relative to model expectations.

The dataset was filtered to include only league or tournament matches (`WEIGHTING` ≥ 0.75) to ensure match seriousness and rating reliability. Players with at least two matches on the same calendar day were identified, and their matches were sorted chronologically. For each qualifying player-day, we retained:

- The **first match of the day**, representing the baseline performance.
- The **last match of the day**, capturing any fatigue-related effects.

The last match is simply the final qualifying match the player played that day (which may be their second match, or one of many). The change in performance deviation between the first and last match was computed:

$$\Delta\text{Deviation} = \text{Deviation}_{\text{last}} - \text{Deviation}_{\text{first}} \tag{2}$$

A Wilcoxon signed-rank test was used to assess whether there was a statistically significant shift in performance deviation between the first and last matches. This non-parametric test was selected due to the non-normality of the distribution and the paired nature of the data. The analysis included over 237,000 player-day pairs, making it suitable for robust statistical inference. Boxplots were used to visualise the shift in deviation across match order.

Effort Modulation in Mismatches

To evaluate whether stronger players systematically adjust their effort in unbalanced matches, we analysed the relationship between pre-match rating disparity and match outcome dominance. In particular, we sought to identify whether players underperform in matches where their skill advantage is large enough to affect motivation or intensity (see Figure 2 for details of the pipeline).

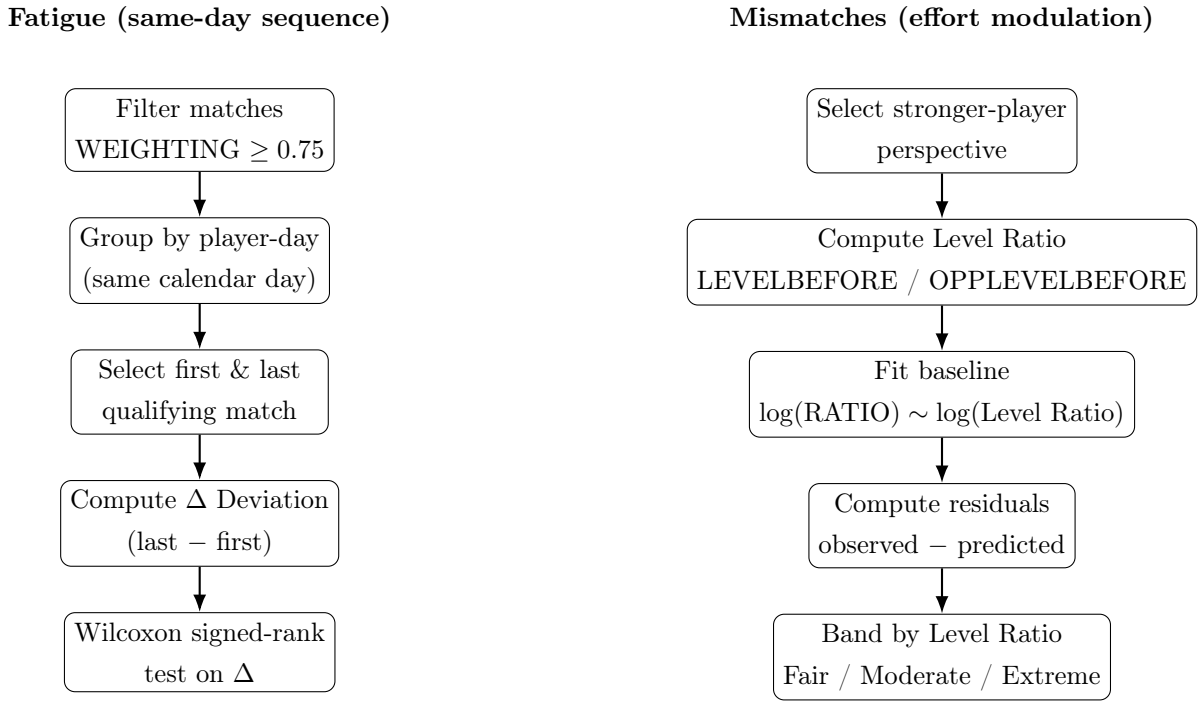

FIGURE 2: Study design and analysis pipeline for the fatigue and mismatch analyses.

Each match was analysed from the perspective of the stronger player, determined by comparing the pre-match `LEVELBEFORE` ratings. The performance outcome was measured using `RATIO`, which quantifies the match dominance based on both games and points won. This metric was log-transformed to stabilise variance and approximate linearity. Similarly, the level difference between players was expressed as the log of the ratio of their pre-match levels.

A simple linear regression was fitted to model expected dominance as a function of the level disparity:

$$\log(\text{RATIO}) = \beta_0 + \beta_1 \cdot \log(\text{Level Ratio}) + \epsilon \tag{3}$$

The fitted model served as a baseline for expected outcomes given the rating difference. To identify possible behavioural effects, we computed the residuals from this model, which were defined as the difference between the observed and predicted log-transformed `RATIO`:

$$\text{Residual} = \log(\text{RATIO}_{\text{observed}}) - \log(\text{RATIO}_{\text{predicted}}) \tag{4}$$

Residuals quantify whether a player outperformed or underperformed relative to expectations derived from the rating model. We then grouped matches into three level ratio bands using thresholds aligned with the SquashLevels rating system (*About the SquashLevels System*, no date):

- **Fair**: Level ratio 1.0–1.5
- **Moderate Mismatch**: Level ratio 1.5–3.0
- **Extreme Mismatch**: Level ratio >3.0

These bands represent increasing pre-match disparity between the stronger and weaker player. A level ratio close to 1 indicates a roughly even contest, while larger ratios indicate that the stronger player is expected to dominate more clearly. The extreme mismatch band (ratio > 3) corresponds to matches where the stronger player is more than three times higher-rated, and where SquashLevels applies additional mismatch handling because outcomes may be less informative about true ability (e.g., due to reduced effort or socially moderated scorelines).

The distribution of residuals was examined across these bands. If effort modulation is occurring, one would expect residuals in the extreme mismatch band to be systematically lower—indicating reduced dominance compared to what the model predicts. A negative residual trend in high-gap matches would suggest players coast or manage effort once victory is assured, supporting the hypothesis of behavioural adjustment in mismatches.

Results

Fatigue Effects

To assess whether player performance deteriorates due to fatigue across multiple matches played on the same day, a paired analysis was conducted comparing each player's deviation from expected outcome in their first and last match of the day. The dataset was filtered to include only matches classified as league or tournament matches—those with a `WEIGHTING` value greater than or equal to 0.75. This ensured that only serious and competitively meaningful matches were included, yielding a final sample of 237,114 player-day pairs where each player participated in at least two qualifying matches.

The outcome variable was the logarithmic deviation from expected performance, calculated as:

$$\text{Deviation} = \log(\texttt{RATIO}) - \log\left(\frac{\texttt{LEVELBEFORE}}{\texttt{OPPLEVELBEFORE}}\right) \tag{5}$$

This deviation measures whether a player overperformed or underperformed relative to their expected result. Positive values indicate better-than-expected performance.

A Wilcoxon signed-rank test was used to assess the difference in deviation between the first and last match of the day. The result was highly statistically significant, with a test statistic of $Z = 13{,}627{,}299{,}905.5$ and a p-value of 1.18×10^{-37}. Descriptive statistics showed a median delta deviation of approximately $+0.015$, suggesting a slight improvement in performance in later matches.

Figure 3 shows a boxplot of deviation scores, revealing a subtle rightward shift in the distribution for players' last matches. Figure 4 illustrates the distribution of the change in deviation:

$$\Delta\text{Deviation} = \text{Deviation}_{\text{last}} - \text{Deviation}_{\text{first}} \tag{6}$$

This distribution is approximately symmetric and centred just above zero.

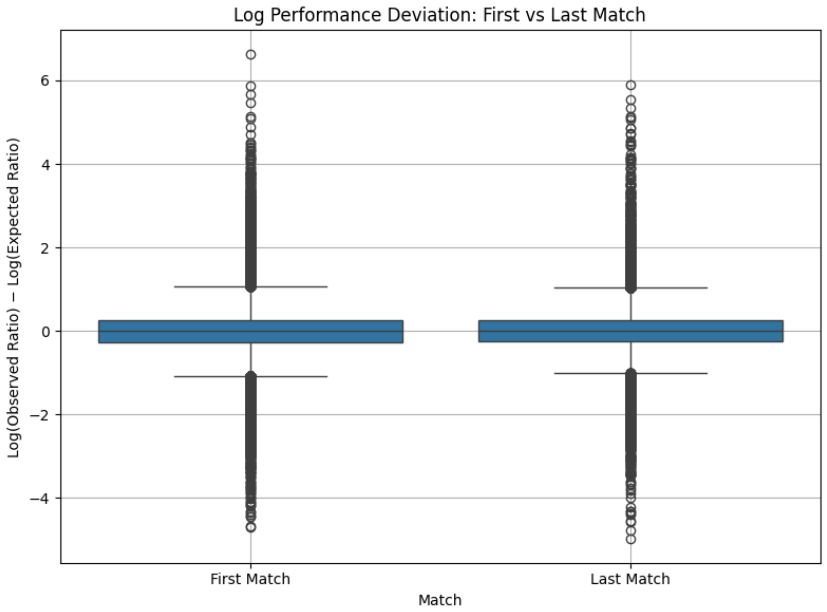

FIGURE 3: Deviation from expected performance in first and last matches of the day. Each point represents the value defined in Equation 5.

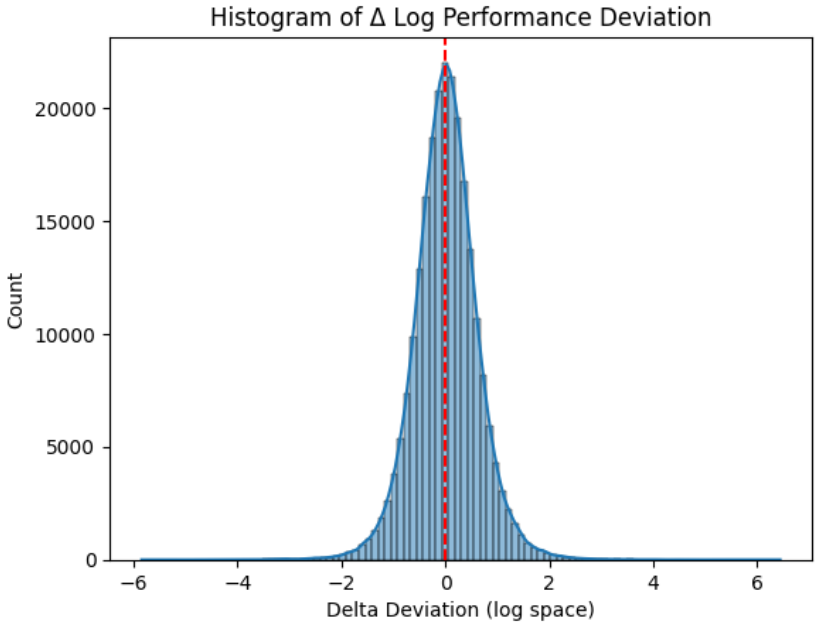

FIGURE 4: Histogram of Δ deviation (Equation 6) across 237,114 player-day pairs. Red vertical line marks the median.

These findings suggest no evidence of fatigue-related underperformance across a day. Instead, players tend to slightly outperform expectations in later matches. This could

reflect warm-up effects, adaptation, or opponent variation. From a rating system perspective, the result indicates that SquashLevels is robust to intra-day performance shifts, and that rating updates are not systematically biased by short-term fatigue effects.

Effort Modulation in Mismatches

To investigate whether stronger players reduce effort in unbalanced matches, a residual-based analysis was conducted using the SquashLevels `RATIO` metric. This captures match dominance from the stronger player's perspective, combining both game and point scores.

A log–log linear regression model was fitted to estimate expected performance based on pre-match rating difference:

$$\log(\text{RATIO}) = \beta_0 + \beta_1 \cdot \log(\text{Level Ratio}) + \epsilon \tag{7}$$

The model was highly statistically significant ($F(1, 3{,}826{,}049) = 1.60 \times 10^6$, $p < 0.001$) with an R^2 of 0.294, indicating that nearly 30% of performance variation could be explained by rating gap. The fitted line was:

$$\log(\text{RATIO}) = 0.058 + 0.641 \cdot \log(\text{Level Ratio}) \tag{8}$$

Residuals from the model were computed as:

$$\text{Residual} = \log(\text{RATIO}_{\text{observed}}) - \log(\text{RATIO}_{\text{predicted}}) \tag{9}$$

Matches were grouped into three bands based on SquashLevels conventions:

- **Fair**: Level Ratio 1.0–1.5
- **Moderate Mismatch**: Level Ratio 1.5–3.0
- **Extreme Mismatch**: Level Ratio > 3.0

Figure 5 displays the distribution of residuals across these bands. In Fair and Moderate Mismatches, residuals were near zero or slightly positive, suggesting performance matched or exceeded model expectations. However, in Extreme Mismatches, the median residual fell to -0.050, indicating that stronger players underperformed relative to prediction.

This supports the hypothesis of *effort modulation*—that stronger players ease off in highly unbalanced matches—consistent with SquashLevels' approach of reducing rating adjustment in these cases.

Figure 6 provides further context, plotting the fitted regression line over observed data. The plot shows a strong central trend, but also large variability, especially for high level ratios.

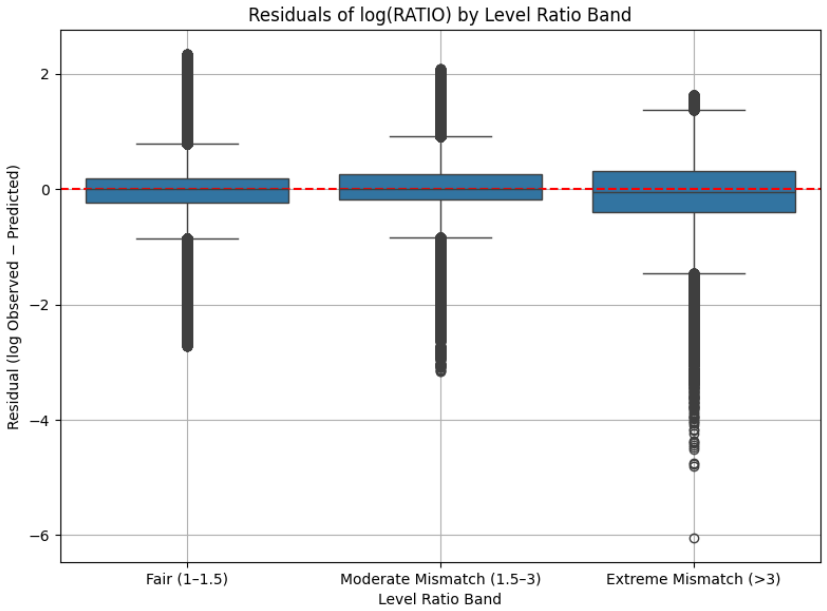

FIGURE 5: Residuals from Equation 9 (Equation **??**) grouped by level ratio band. The red dashed line indicates zero (perfect prediction). Underperformance appears in extreme mismatch cases.

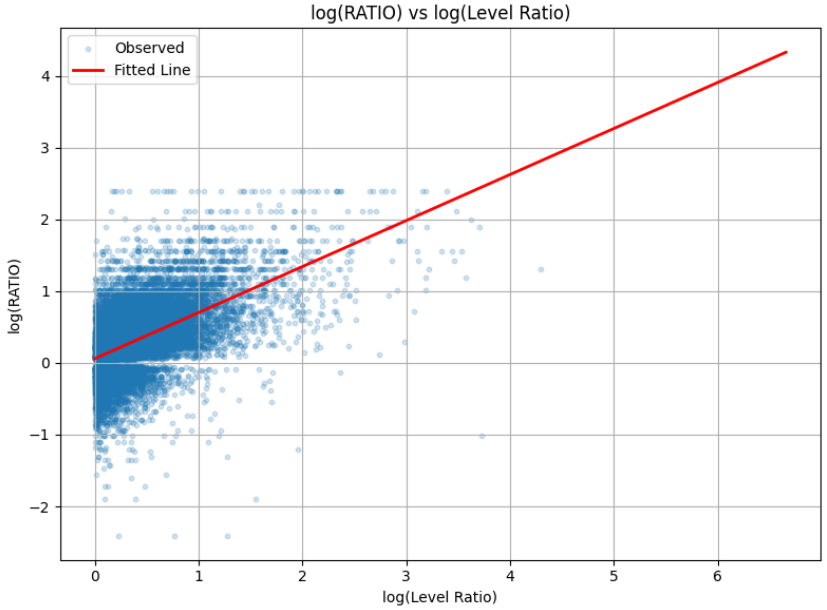

FIGURE 6: Scatterplot of log(RATIO) vs log(Level Ratio) for stronger players. Red line is the fitted model (Equation 8).

Discussion

This chapter returns to the project's central claim: that some observed match outcomes may deviate systematically from SquashLevels predictions due to *behaviour*, rather than genuine changes in underlying playing ability. The purpose here is not to restate the results in detail, but to interpret what the two analyses imply about (i) the plausibility of behavioural anomalies in this setting, (ii) how the two mechanisms relate to each other, and (iii) what the patterns suggest about SquashLevels' modelling assumptions.

Fatigue Effects

The fatigue analysis was motivated by a straightforward mechanism: where a player competes multiple times in one day, accumulated physical and cognitive load could reduce later performance. If fatigue were a systematic behavioural anomaly, the final match of the day would be expected to deviate negatively from the SquashLevels expectation more often than the first match.

The observed pattern did not align with this expectation. Rather than interpreting this as proof that fatigue does not exist in squash, the more defensible interpretation is that fatigue is not a dominant driver of rating-relative deviations under the conditions captured by this dataset and filtering. Several contextual explanations remain consistent with the findings. First, warm-up effects can plausibly offset or dominate fatigue at the day-level ordering used here: players may take time to reach full movement efficiency, timing, and decision quality. Second, selection effects are likely: the subset of players who play multiple competitively weighted matches in a day may be fitter, more experienced, or engaged in formats where match load is expected and managed. Third, the dataset does not encode key mediators of fatigue (e.g., rest time, match duration, cumulative points played), so any fatigue signal may be diluted when comparing only first versus last matches.

An important implication is methodological: day-level match ordering is a coarse proxy for fatigue. The lack of a negative shift therefore suggests that, within the current

measurement resolution, SquashLevels' assumption of stable within-day performance is not obviously violated in a way that would systematically bias ratings. In other words, the analysis provides weak support for fatigue as a behavioural anomaly of practical relevance in this data, while leaving open the possibility that more granular features could reveal conditional fatigue effects.

Effort Modulation in Mismatches

Effort modulation was hypothesised to occur when a match is highly unbalanced. In such contexts the stronger player may reduce intensity once the outcome is effectively secure, producing scorelines that are less dominant than would be expected purely from the rating gap. Unlike fatigue, this mechanism is inherently contextual: it should be most visible in large skill disparities and may not appear at all in balanced matches.

The residual-based analysis supports this behavioural mechanism most clearly in the *extreme mismatch* regime. Crucially, this inference depends on comparing observed dominance to an explicit baseline expectation. The regression residual framing therefore strengthens the behavioural interpretation: it isolates the question of whether performance is systematically *lower than predicted*, not merely whether stronger players tend to win (which is tautological in a rating system). The negative shift in residuals for extreme mismatches is consistent with reduced intensity, pacing, or social norm effects (e.g., avoiding excessive blowouts), and it aligns with SquashLevels' documented reasoning for dampening adjustments in highly unequal contests (SquashLevels, no date).

This result also highlights a key theoretical point: behavioural anomalies do not necessarily undermine a rating system; they often motivate the system's practical heuristics. In this case, the system's mismatch handling can be interpreted as a behavioural correction rather than a purely statistical device. The analysis therefore offers empirical coherence between algorithm design and observed match behaviour.

Integrating the two mechanisms

A useful way to integrate both findings is to distinguish between load-driven and context-driven behavioural effects. Fatigue is load-driven and depends on physiological depletion and recovery, which are not directly observed here beyond match ordering. Effort modulation is context-driven and depends on incentives, perceived challenge, and the marginal value of additional points or games. The results suggest that, in this dataset, context-driven effects are more visible at scale than load-driven effects.

This contrast also clarifies what "behavioural anomaly" means in practice for SquashLevels: not every plausible behavioural mechanism generates a detectable systematic deviation in rating-relative outcomes, but certain match contexts (notably extreme mismatches) produce consistent departures from baseline expectations. This distinction provides a coherent bridge into the next chapter, which focuses on what should be concluded, what can be recommended, and what limitations constrain those recommendations.

Conclusions and Further Work

Conclusions

Readers are already familiar with the two empirical patterns established in the Results chapter and interpreted in the Discussion: (i) there was no evidence of a practically meaningful fatigue-related decline using a first-versus-last match comparison within competitively weighted fixtures, and (ii) there was clear evidence that extreme mismatches produced systematically lower-than-expected dominance by the stronger player when evaluated relative to a baseline predictive model. This chapter consolidates what those patterns mean for the project's unifying thesis and sets out the practical and theoretical value of the work, while also being explicit about limitations and priorities for future research.

The unifying thesis proposed that behavioural factors can create systematic deviations between observed match outcomes and those predicted by the SquashLevels model, without reflecting true changes in underlying playing ability. Based on the evidence in this dissertation, the thesis was partially accepted: behavioural anomalies were not universal across mechanisms, but they were observable and practically relevant in specific match contexts.

Two main conclusions followed. First, within competitively weighted matches and the day-level comparison used here, fatigue did not appear to introduce a systematic downward bias in rating-relative performance. The fatigue test used $n = 237{,}114$ paired player-day observations and returned $p = 1.18 \times 10^{-37}$, but the median shift in deviation was small and positive (approximately $+0.015$). Given the very large sample size, this was best interpreted as a statistically detectable but practically modest shift, consistent with warm-up or selection effects rather than fatigue-driven underperformance. Second, extreme mismatches showed consistent underperformance by the stronger player relative to baseline model expectations, consistent with effort modulation. The log–log regression baseline achieved $R^2 \approx 0.294$ over $n = 3{,}826{,}051$ matches, and the extreme mismatch band showed a negative median residual (approximately -0.050), whereas fair

and moderate bands were near zero or slightly positive. This provided clear evidence of a context-driven behavioural effect concentrated in high-disparity contests.

These findings led directly to practical recommendations for SquashLevels and similar ranking systems. Most importantly, mismatch dampening should be retained and viewed as a behavioural correction: the residual pattern in extreme mismatches provided empirical justification for reduced sensitivity in rating updates when disparities are very large, treating such matches as lower-information events and limiting distortion from low-intensity scorelines. By contrast, there was no evidence in this work to justify a generic same-day fatigue adjustment, and implementing one without richer scheduling features could introduce bias where none was demonstrated. More generally, behavioural corrections appeared most defensible when targeted to match contexts where the deviation signal was strong and consistent, rather than applied uniformly.

Beyond practical guidance, the project contributed a clear modelling lesson: behavioural anomalies in an Elo-like ranking setting were more readily detected when tied to incentive structure and match context (effort modulation) than when they depended on unobserved mediators such as recovery time and physiological load (fatigue). This supports the broader principle that when behavioural mechanisms are hypothesised, the strongest empirical tests compare outcomes to an explicit baseline expectation and examine systematic residual structure rather than relying only on raw outcome differences.

Limitations and Further Work

Several limitations constrain interpretation and indicate where further work would be most valuable. For fatigue, "first vs last match of day" was a coarse proxy and did not account for recovery time, match duration, cumulative points played, or multi-day load. As a result, this dissertation cannot claim that fatigue effects are absent in squash; rather, it found no evidence of fatigue-driven underperformance under this proxy and filtering. A direct extension would incorporate richer fatigue proxies, such as time gaps between matches, duration proxies (e.g. points played), and within-tournament sequences spanning multiple days.

For mismatch analysis, the baseline regression model used a single predictor (log level ratio). While this simplicity supported transparency, residual patterns may partly reflect omitted contextual variables (e.g. scoring format, match weighting, and confidence). Future work should extend the baseline model with covariates such as `WEIGHTING`, confidence measures, and score type, then reassess residual shifts by mismatch severity. In addition, given the very large samples typical in this dataset, future reporting should

place greater emphasis on practical effect sizes and uncertainty (e.g. bootstrap confidence intervals for median shifts) alongside significance tests.

Finally, a useful applied extension would be algorithm stress-testing: simulate rating trajectories with and without mismatch dampening to quantify downstream impacts on ranking stability and predictive accuracy. This would translate the behavioural findings into clearer guidance on whether current heuristics are near-optimal or could be improved.

In summary, this dissertation showed that behavioural anomalies in SquashLevels were context-dependent. Using simple, reproducible analyses, fatigue effects were not found to produce systematic underperformance in competitively weighted same-day match sequences, while extreme mismatches produced clear and systematic underperformance relative to baseline expectations. The key value of the thesis is that it both validates a major behavioural assumption already embedded in the SquashLevels algorithm (effective effort in mismatches) and provides evidence against introducing an additional generic correction (same-day fatigue) without richer data.

Self-evaluation of Project Outcome

This project required me to work in a domain that was new to me: sports ranking systems and behavioural effects in performance data. Over the course of the dissertation, I developed a strong understanding of how Elo-like systems operate in practice and how SquashLevels adapts these ideas for squash. I also became more confident in reading academic literature, particularly in distinguishing between papers that propose new ranking models and those that evaluate existing systems and behavioural mechanisms. This improved my ability to extract defensible assumptions and position my work within published research.

A key strength of the project was that I maintained a scientific approach even when results did not support my initial expectations. In particular, the fatigue analysis did not show the expected underperformance, but I treated this as an informative outcome rather than forcing a narrative. I followed the scientific method as taught in the Research Methods in Data Science module: defining a testable hypothesis, selecting appropriate metrics, using statistical tests aligned with distributional assumptions, and interpreting results critically. I also kept structured notes throughout, which made it easier to resume work efficiently after interruptions and ensured consistency between the methodology, results, and final write-up.

In terms of process, I applied project management principles from the same module to organise tasks, plan the structure of the dissertation early, and track progress against milestones. This helped me stay focused and avoid scope creep, especially once I decided to narrow the dissertation to two core behavioural mechanisms.

If I were to repeat the project, there are two main areas I would improve. First, although I used project management methods, I struggled to maintain the original timeline when project requirements changed or unfavourable external factors emerged. In future, I would liaise more proactively with my supervisor to re-plan deliverables earlier and reduce last-minute restructuring. Second, I found it challenging to judge the appropriate depth for each chapter. Next time, I would request access to high-quality past dissertations

(where available) to calibrate expectations and strengthen my writing plan from the start.

Appendix A

GenAI Usage Statement

Generative AI tools were used to support drafting and language editing, and to assist with producing and refining analysis code. All outputs were reviewed, edited, and verified by the author. No personally identifiable information was provided to any GenAI tool.

Bibliography

About the SquashLevels System (2026). en. URL: https://app.squashlevels.com/doc.php?doc=about_squash_levels.htm (Accessed: Feb. 6, 2026).

Csató, L. (2021). "Topics in Tournament Ranking". en. In: *Tournament Design: How Operations Research Can Improve Sports Rules*. Ed. by L. Csató. Cham: Springer International Publishing, 1–31. ISBN: 978-3-030-59844-0. https://doi.org/10.1007/978-3-030-59844-0_1.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. en. Google-Books-ID: 8pMnAQAAMAAJ. Arco Pub. ISBN: 978-0-668-04721-0.

Girard, O. and G. P. Millet (2009) "Neuromuscular Fatigue in Racquet Sports". *Physical Medicine and Rehabilitation Clinics of North America*, 20(1), pp. 161–173. Available at: https://doi.org/10.1016/j.pmr.200

Glickman, M. (2011). "The Glicko system". In: URL: https://www.semanticscholar.org/paper/The-Glicko-system-Glickman/e27b5e5642305afdfb93420fe956827ec035da15 (Accessed: Feb. 18, 2026).

Herbrich, R., T. Minka, and T. Graepel (2006). "TrueSkill™ : A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2006/hash/f44ee263952e65b3610b8ba51229d1f9-Abstract.html (Accessed: Feb. 18, 2026).

Holdaway, C. and E. Vul (May 2021). *Risk-taking in adversarial games: What can 1 billion online chess games tell us?* https://doi.org/10.31234/osf.io/vgpdj.

Irons, D. J., S. Buckley, and T. Paulden (2014) "Developing an improved tennis ranking system". *Journal of Quantitative Analysis in Sports*, 10(2), pp. 109–118. Available at: https://doi.org/10.1515/jqas-2013-0101.

Kelley, J. (2024) "Using simulations to compare the current Davis Cup ranking system to Elo". *PLOS ONE*, 19(2), pp. e0298188. Available at: https://doi.org/10.1371/journal.pone.0298188.

Kovalchik, S. (2020) "Extension of the Elo rating system to margin of victory". *International Journal of Forecasting*, 36(4), pp. 1329–1341. Available at: https://doi.org/10.1016/j.ijforecast.2020.01.006.

Krumer, A., R. Megidish, and A. Sela (2023) "Strategic manipulations in round-robin tournaments". *Mathematical Social Sciences*, 122, pp. 50–57. Available at: https://doi.org/10.1016/j.mathsocsci.2023.0

Nieken, P. and M. Stegh (Jan. 2010). *Incentive Effects in Asymmetric Tournaments Empirical Evidence from the German Hockey League*. eng. doc-type:workingPaper. Volume: 305. https://doi.org/10.5282/ubm/epub.13249.

Vaziri, B., S. Dabadghao, Y. Yih, and T. L. Morin (2018) "Properties of sports ranking methods". *Journal of the Operational Research Society*, 69(5), pp. 776–787. Available at: https://doi.org/10.1057/s41274-017-0266-8.