Objective Algorithms, Context-Dependent Performance, and the Incorporation of Bias: Evidence from Chess*

Guangli Zhang† Michael Podgursky‡ Michelle Wickman§

July 6, 2026

Abstract

Objective performance metrics are often treated as neutral measures of talent. This paper studies a setting in which the measurement rule is objective, explicit, and mechanical, but the performance signal it incorporates may be context-dependent. Using rated chess games, we examine whether female players' high-stakes performance response is attenuated when the opponent is male, and whether these local score wedges are mechanically capitalized into future ratings. We exploit the U.S. Chess Federation bonus-point rule, which creates sharp event-level incentives for unusually strong performance. Among female players near the bonus threshold, we compare game outcomes against male and female opponents. Female players exhibit a positive near-threshold score response against female opponents, but this response is substantially attenuated against male opponents. We interpret this interaction as a score-unit wedge in high-stakes responsiveness and map it into the rating update rule. Feeding the estimated wedge through the rating algorithm, we construct counterfactual rating histories that net out the local attenuation. The replay implies meaningful rating gaps that are right-skewed and strongly exposure-driven: players with more male-opponent high-stakes games accumulate larger gaps in measured ability. We then compare alternative updating schedules and show that faster learning can also imply faster incorporation of context-dependent performance wedges, generating an equity–efficiency frontier in talent measurement. The results highlight a general point: even when measurement rules are objective, explicit, and mechanical, measured talent can be biased when the performance signals entering the rule are context-dependent.

Keywords: talent measurement, ratings, performance evaluation, gender, chess, incentives, algorithmic measurement.
JEL codes: J16, J24, J71, D83, C23, C55.

*All errors are our own. Preliminary draft, please do not circulate without permission.

†Washington University in St. Louis, Center for Social Development (email: guangli@wustl.edu)
‡The University of Missouri, Department of Economics (email: podgurskym@missouri.edu)
§Saint Louis University, Sinquefield Center for Applied Economic Research (email: michelle.wickman@slu.edu)

1 Introduction

Many markets rely on ratings, scores, rankings, and algorithmic assessments to quantify talent. These metrics are often treated as objective because the rules that generate them are transparent, standardized, and mechanically applied. Whether in test scores, promotion metrics, or platform reputations, the procedural logic appears neutral: everyone is subject to the same formula, which updates strictly based on observed outcomes. Yet procedural objectivity does not guarantee measurement neutrality. The inputs to these rules may themselves be shaped by context-dependent frictions. If individuals systematically face environments that suppress their performance, a rule that appears neutral will mechanically convert those situational disadvantages into durable assessments of ability. A measurement system need not explicitly discriminate to generate inequality; it only needs to treat context-constrained performance as a pure reflection of underlying talent.

This paper studies this incorporation problem in chess. We distinguish between two objects that are often conflated. The first is a *flow distortion*: a current-game performance wedge that arises in a particular setting. The second is a *stock distortion*: the cumulative rating gap that results when prior flow distortions are capitalized into future ratings. This distinction matters because ratings are not only summaries of past performance. They are also state variables that shape future expected scores, tournament classifications, and the interpretation of subsequent outcomes. The central question is therefore not only whether performance is context-dependent, but whether an objective rating rule converts context-dependent performance into a durable measure of ability.

We formalize this idea in a simple framework that separates latent ability, game-specific performance, and the rating system's mechanical interpretation of performance. The framework captures a basic tension: high-stakes games increase the value of effort, attention, and error checking, but the same environments may also raise the cognitive cost of supplying those inputs. As a result, incentives need not translate mechanically into better performance. They may instead generate attenuated performance responses in contexts where cognitive load is especially high. This framework guides the empirical analysis in two steps. First, we use variation around bonus thresholds to estimate whether female players' high-stakes score response is attenuated when the opponent is male. Second, because chess ratings mechanically update from realized scores, we replay the rating rule to measure how much of this current-game wedge is incorporated into future rating-based measured ability.

Chess provides a useful testing ground for this problem. The setting has the core features of many talent-measurement systems, but in a more transparent and empirically tractable form. Performance is repeatedly observed, game outcomes are objectively recorded, and ratings are updated by explicit formulas rather than by opaque or discretionary evaluations. At the same time, chess is not a frictionless measure of ability. Players make decisions under time pressure, strategic uncertainty, and varying stakes, so realized performance can depend on the context in which ability is expressed. This combination is precisely what makes the setting valuable: the measurement rule is clean, but the performance input to that rule may be context dependent. Chess also allows us to observe the full sequence linking current performance to future measured ability. We see players'

ratings, opponents, game outcomes, and rating-rule parameters over time, which lets us distinguish local performance effects from cumulative rating effects. Finally, because rating systems differ in how aggressively they update beliefs from new outcomes, the setting allows us to ask a broader design question: how alternative updating rules would translate the same current-game distortion into different long-run measures of talent.

Using this setting, we find that female players' high-stakes performance response is substantially attenuated when the opponent is male. In the preferred bonus-threshold specification, female players exhibit a positive near-bonus score response against female opponents, but this response is almost entirely muted against male opponents. The interaction estimate implies a score-unit attenuation wedge of about 0.0098 points, or roughly 85 percent of the baseline near-bonus response.

The rating replay shows why this local wedge matters. We add the estimated missing score back to affected games and mechanically replay the rating rule. This produces a score-equivalent measure of total distortion: the direct missing score in the current affected game plus the expected-score loss already embedded in the player's rating from prior affected games. Roughly four-fifths of this total distortion is already absorbed into ratings before the current game begins. The implication is that observed ratings should not be interpreted only as controls for ability; they are also cumulative outcomes of past context-dependent performance.

We then use the estimated score wedge to study rating-system design more generally. Holding the local attenuation estimate fixed, we ask how alternative updating rules would translate the same distorted performance signal into long-run ratings. In a stylized simulation with known latent ability, we compare the observed USCF rule to a family of alternative K schedules that vary how aggressively ratings respond to new information. For each rule, we compute two objects: an incorporation-bias measure, defined as the mean counterfactual rating gain from adding back the attenuation wedge, and a learning-speed measure, defined as the number of games required for ratings to converge near latent ability. This exercise traces an efficiency–equity frontier. Faster-updating rules learn more quickly from performance data, but they also capitalize context-dependent distortions into ratings more aggressively. The observed USCF rule lies on the fast-learning side of this trade-off, but it is neither neutral nor generally frontier-efficient. The implication is not that rating systems should simply update more slowly. Rather, the same parameter that governs responsiveness to new information also governs the speed at which distorted performance signals become embedded in measured talent.

Our paper contributes to a growing literature that uses chess as a high-frequency field setting to study gender, competition, and performance (Brancaccio & Gobet, 2023; Kisida et al., 2026).[1] For our purposes, a central feature of this setting is that chess ratings are not merely descriptive statistics; they are the institutionalized measure of competitive ability. Prior work studies gender gaps in

[1]Chess is especially useful for studying performance measurement because outcomes are objectively scored, repeated over time, and summarized through public rating systems (Elo, 1978; Glickman, 2001). It therefore provides a clean laboratory for studying a broader class of dynamic talent-measurement systems. Elo and Elo-type rating systems have also been adapted to many other competitive settings, including association football (Hvattum & Arntzen, 2010), tennis (S. A. Kovalchik, 2016; S. Kovalchik, 2020), Go (Belgian Go Federation, 2024), online games, and other sports (Düring et al., 2022).

chess participation and ratings (Chabris & Glickman, 2006; Bilalić et al., 2009), gender differences in strategic behavior (Gerdes & Gränsmark, 2010), and context-dependent performance in mixed-gender competition. In closely related work, De Sousa & Hollard (2023) show how small micro-level gender differences in chess games can aggregate into larger macro-level gaps through effort, beliefs, human-capital investment, and quitting. Their paper highlights an important behavioral accumulation problem: even modest performance differences in repeated competitive environments can have large long-run consequences.

Closest to our empirical setting, Backus et al. (2023) show that women make more mistakes when playing against men, while men do not exhibit comparable declines in quality of play against female opponents. Related work debates the role of stereotype threat in chess. Stafford (2018) finds that female chess players outperform rating-based expectations when playing men, while Smerdon et al. (2020) argue that this conclusion changes once the specification accounts more flexibly for player heterogeneity, yielding evidence consistent with stereotype-threat effects.[2] Importantly for our purposes, Smerdon et al. (2020) also highlight a deeper measurement problem: because chess ratings are updated using past game outcomes, rating-based expectations may already incorporate the same gendered performance effects they are later used to benchmark. Our paper takes this observation as a starting point and studies a distinct accumulation channel. Rather than focusing on behavioral responses, human-capital investment, or attrition, as in De Sousa & Hollard (2023), we examine how the rating rule itself mechanically capitalizes context-dependent performance into future rating-based measured ability. We formalize this incorporation channel, identify a local high-stakes opponent-gender attenuation wedge using the USCF bonus-threshold rule, and quantify how that wedge becomes embedded in future ratings through a formally gender-blind updating rule.

Our paper also contributes to the economics literature on gender, competition, and context-dependent performance (Croson & Gneezy, 2009; Niederle & Vesterlund, 2011). This literature shows that gender gaps can emerge in competitive performance, tournament entry, self-evaluation, self-promotion, and career-relevant choices (Gneezy et al., 2003; Niederle & Vesterlund, 2007; Buser et al., 2014; Exley & Kessler, 2022). Recent work on cognitive uncertainty provides a particularly useful behavioral foundation for our model. Enke & Graeber (2023) show that individuals differ in their subjective uncertainty about the optimal action, that such uncertainty predicts systematic decision distortions, and that women report higher cognitive uncertainty across several decision domains. In our setting, this motivates a simple but important distinction between latent ability and realized performance: even when ability is fixed, observed play may vary with the decision environment if some players experience greater uncertainty, hesitation, or sensitivity to context. We differ from much of the gender-and-competition literature by studying a naturally occurring, high-stakes environment in which realized performance is not only observed but also mechanically incorporated into future evaluations. The central question is therefore not only whether performance is context-dependent, but whether such dependence becomes embedded in durable measures

[2]Experimental evidence also suggests that gender stereotypes can affect chess performance (Maass et al., 2008; Rothgerber & Wolsiefer, 2014).

of ability.

Our paper also contributes to personnel economics and the literature on incentives and performance measurement (Prendergast, 1999). A central lesson of this literature is that incentives depend critically on what is measured, and that imperfect performance measures can distort behavior (Holmström, 1979; Holmström & Milgrom, 1991; Baker, 1992; Lazear & Rosen, 1981). Work on nonlinear incentives further shows that convexities, thresholds, and discontinuities in compensation and evaluation rules can generate sharp behavioral responses, including changes in effort, timing, reporting, and risk-taking (Healy, 1985; Oyer, 1998; Larkin, 2014; Chevalier & Ellison, 1997). Other work studies how organizations use measured performance to allocate future opportunities. Benson et al. (2019) show that firms promote high-performing sales workers even when current sales performance is an imperfect predictor of managerial performance, highlighting a tension between using performance metrics to provide incentives and using them to measure suitability for future roles. Closest in spirit, Brown (2011) shows that competing against superstars can reduce rivals' performance in professional golf. Related evidence from professional sports shows that social context can also directly affect realized performance: Caselli et al. (2023) find that African football players in Italy improved when COVID restrictions removed stadium crowds, especially in settings with more prior racist behavior. Like these papers, we study how competitive context and performance measurement interact. Our contribution is to extend this logic from worker behavior and organizational assignment to the measurement of talent itself. In our setting, ratings are not discretionary evaluations and are not subject to manager favoritism; they are produced by a transparent algorithm using objective win-loss outcomes. Yet even such a system can fail to isolate context-free ability if the input signal is context-dependent. The performance-measurement problem therefore does not disappear when evaluation becomes objective. Instead, an objective algorithm can convert context-contaminated performance signals into persistent measures of talent.

Our paper is also related to the literature on discrimination, statistical discrimination, and systemic bias (Becker, 1957; Phelps, 1972; Arrow, 1973; Aigner & Cain, 1977; Coate & Loury, 1993). Much of this literature studies biased evaluators, biased beliefs, or differential treatment by decision-makers. For example, audit and resume-rating designs identify how employers respond to candidate characteristics, either through field experiments or incentivized evaluations that preserve incentives without deception (Bertrand & Mullainathan, 2004; Kessler et al., 2019). Related work in sports and workplace settings shows that unequal treatment can distort measured productivity itself. Parsons et al. (2011) show that racial bias in baseball umpires' strike calls causes minority pitchers' productivity to be systematically underestimated, implying that even a race-neutral pay-for-performance rule may fail to reward equally productive workers equally. Other work shows that unequal treatment can affect actual performance and thereby reinforce future evaluations (Glover et al., 2017; Bohren et al., 2019). Our setting differs because the rating formula does not directly use gender and does not require evaluator discretion. Nevertheless, bias can enter upstream through realized performance and persist downstream through mechanical updating. This connects our paper to recent work on systemic discrimination, which emphasizes that group disparities can

arise not only from the direct use of group identity in the focal decision, but also from upstream processes that shape the signals, qualifications, or histories on which nominally neutral decisions rely (Bohren et al., 2025). We provide a transparent empirical example of this mechanism in an algorithmic evaluation system: a gender-blind updating rule can mechanically encode unequal histories when the performance outcomes used for updating are generated under context-dependent conditions.

Finally, our paper speaks to the economics of algorithms, machine learning, and prediction systems (Mullainathan & Spiess, 2017; Athey & Imbens, 2019; Agrawal et al., 2018, 2019). Recent work emphasizes that algorithmic systems may reproduce or amplify bias when they rely on biased labels, proxy outcomes, or historically distorted data (Kleinberg et al., 2018, 2019; Obermeyer et al., 2019; Cowgill et al., 2020). Chess ratings provide a particularly clean environment for studying this concern because the algorithm is transparent, the inputs are observed, and the updating rule is known. Unlike many complex prediction systems, the rating rule is simple enough to make the incorporation channel explicit. Our paper therefore offers a tractable economic setting for studying a general problem in algorithmic measurement: prediction systems can accurately summarize realized outcomes while still embedding inequities in the process that generated those outcomes. In this sense, we use chess to study the tradeoff between learning quickly from performance data and avoiding the mechanical capitalization of context-dependent performance distortions into durable measures of talent.

The rest of the paper is organized as follows. Section 2 describes the USCF rating system, the bonus-point rule, and the institutional parameters that generate the high-stakes states used in the empirical design. Section 3 develops a simple framework linking incentives, context-dependent performance costs, and the mechanical incorporation of realized performance into measured ability. Section 4 describes the game-level panel and the female-focal analytical sample. Section 5 presents the empirical design for estimating context-dependent high-stakes performance responses. Section 6 reports the reduced-form evidence on male-opponent attenuation. Section 7 uses the estimated score wedge to replay the rating rule and measure how local flow distortions become embedded in the stock of ratings. Section 8 studies policy-design implications by comparing alternative update schedules and tracing an equity–efficiency frontier. Section 9 concludes.

2 Institutional Background: USCF Ratings and Bonus Incentives

Chess provides an unusually clean setting for studying talent measurement. Game outcomes are objective, repeated, and publicly recorded. A player either wins, draws, or loses, generating a score $S_{ig} \in \{1, 0.5, 0\}$. These game-level outcomes are then transformed into ratings through an explicit institutional rule. Ratings are consequential because they summarize past performance, affect future expected scores, enter tournament sectioning and classification, and serve as the public measure of chess ability.

The empirical setting is the United States Chess Federation (USCF) rating system. The unit

of updating is an event rather than an individual game. For player i in event e, let S_{ie} denote the player's total event score and let E_{ie} denote the total expected score implied by the player's rating and the ratings of the opponents faced in the event. At the game level, the expected score is Elo-style:

$$E_{ig} = \frac{1}{1 + 10^{(R_{jg} - R_{ig})/400}}, \tag{1}$$

and the event-level expected score is $E_{ie} = \sum_{g \in e} E_{ig}$. The difference $S_{ie} - E_{ie}$ is therefore the player's event-level over-performance relative to the rating system's expectation.

For established players, the event-level rating update can be written as

$$\Delta R_{ie} = K_{ie}(S_{ie} - E_{ie}) + \max\left\{0,\ K_{ie}(S_{ie} - E_{ie}) - B_t\sqrt{m'_{ie}}\right\}, \tag{2}$$

where K_{ie} is the rating sensitivity parameter, B_t is the bonus threshold parameter in force at time t, m_{ie} is the number of games played in the event, and $m'_{ie} = \max\{m_{ie}, 4\}$. The first term is the ordinary rating update. The second term is the bonus adjustment. Bonus points are extra rating points awarded only when event-level over-performance exceeds the threshold $B_t\sqrt{m'_{ie}}$.

The K parameter determines how strongly realized performance is converted into rating changes. In the post-2001 USCF system, K is a sliding rather than fixed parameter. A useful representation is

$$K_{ie} = \frac{800}{N_{0,ie} + m_{ie}}, \tag{3}$$

where $N_{0,ie}$ is the player's effective number of prior games. Players with fewer effective games receive larger updates, while more established players receive smaller updates. Thus, K is both a statistical learning-rate parameter and an institutional rule governing how quickly observed performance becomes measured ability.

The relevant institutional parameters, including the bonus-threshold values and the effective-games rules used to compute K_{ie}, are summarized in Table 1.

The bonus rule creates the high-stakes states used in the empirical design. Bonus points begin when

$$K_{ie}(S_{ie} - E_{ie}) > B_t\sqrt{m'_{ie}}. \tag{4}$$

Equivalently, the player must exceed expected score by more than

$$S_{ie} - E_{ie} > \frac{B_t\sqrt{m'_{ie}}}{K_{ie}}. \tag{5}$$

This expression makes clear how institutional parameters affect the difficulty of reaching the bonus threshold. A higher B_t makes the threshold harder to reach. A higher K_{ie} makes the threshold easier to reach in score units, because each unexpected point generates more rating movement. Longer events require more absolute over-performance through the $\sqrt{m'_{ie}}$ term.

This structure generates rich within-event variation. At any point in an event, a player's accumulated score and expected score determine the distance to the bonus threshold. A player

just below the threshold faces a larger marginal rating return to an additional result than a nearby player farther away from the threshold. The empirical design uses this institutional feature to define bonus-relevant high-stakes states and asks whether players' performance responses differ depending on opponent gender.

The rating system also changed over time. The bonus threshold parameter B_t was adjusted several times, and the effective-games formula used to determine K was revised in 2013. These changes are useful institutionally because they underscore that K and B_t are policy parameters rather than behavioral primitives. Table 1 reports the parameter values used to reconstruct the rule in each period, and Figure 1 summarizes the bonus-threshold changes relevant for coding the institutional rule over time. In the empirical analysis, we use the parameter values corresponding to each event date and absorb common tournament- and rule-regime variation using event-section-round-score-cell fixed effects.

Several additional institutional features are useful for interpretation. Ratings can be provisional for players with limited rated histories, and the rating formula includes special treatment for very new players and extreme early records. Tournament organizers also use ratings to assign players to sections and classes, so rating changes can affect not only measured ability but also future competitive environments. These features reinforce the broader point that ratings are not passive summaries of performance. They are state variables embedded in an institutional system.

For the purposes of the paper, the key advantage of the setting is that both the signal and the updating rule are observed. We observe objective game outcomes, opponent ratings, player histories, tournament structure, and the formula that maps performance into ratings. This allows us to move from a reduced-form score response to a counterfactual rating replay. The institutional rule tells us how a local score-unit wedge would be converted into rating points, and the panel structure lets us study how repeated exposure can accumulate into measured talent.

3 Conceptual Framework

This section develops a simple framework that links incentive variation, context-specific performance costs, and the mechanical incorporation of realized performance into measured ability. The model is intentionally reduced-form. It does not require separately identifying the primitives of the effort-cost function. Instead, it clarifies why the empirical interaction coefficient is the sufficient statistic for the counterfactual rating replay.

3.1 Incentives and Endogenous Game Performance

We focus on female players. Player i has fixed latent ability A_i. In game g, observed performance is

$$p_{ig} = A_i + \alpha e_{ig} + \varepsilon_{ig}, \qquad \alpha > 0, \tag{6}$$

where $e_{ig} \geq 0$ denotes game-specific effort, attention, calculation depth, and error checking. The shock ε_{ig} captures idiosyncratic performance variation, with

$$\mathbb{E}\left[\varepsilon_{ig} \mid e_{ig}\right] = 0. \tag{7}$$

Let I_{ig} denote the marginal value of performance in game g. In the chess application, I_{ig} is higher when the player is close to the USCF bonus-point threshold, because an additional score point can trigger a discrete increase in the rating reward from the event. More generally, I_{ig} captures any local increase in the stakes attached to performance.

Let M_{ig} be an indicator equal to one when the female focal player faces a male opponent. The player chooses effort to solve

$$\max_{e \geq 0} \quad I_{ig}\mathbb{E}\left[p_{ig} \mid e\right] - C\left(e, \theta M_{ig}\right), \tag{8}$$

where $C(e, \theta M)$ is increasing and convex in effort. The parameter $\theta \geq 0$ is a fixed context-specific cost that turns on in female-against-male games. We interpret θ broadly as a cognitive-tax state: when $M_{ig} = 1$, sustaining concentration, calculation, or error control is more costly.[3] Since

$$\mathbb{E}\left[p_{ig} \mid e\right] = A_i + \alpha e,$$

the choice problem is equivalent to

$$\max_{e \geq 0} \quad \alpha I_{ig} e - C\left(e, \theta M_{ig}\right). \tag{9}$$

Assume

$$C_e(e, \theta M) > 0, \qquad C_{ee}(e, \theta M) > 0.$$

The Kuhn–Tucker conditions imply

$$\alpha I \leq C_e(0, \theta M) \quad \Rightarrow \quad e^*(I, M) = 0,$$

and

$$\alpha I = C_e\left(e^*(I, M), \theta M\right) \quad \text{if } e^*(I, M) > 0. \tag{10}$$

Below, we focus on the interior range relevant for the empirical variation.

3.2 Context-Specific Responses to High Stakes

The model highlights a simple trade-off. Higher incentives (I) raise the marginal benefit of effort. At the same time, in female-against-male games, the fixed contextual cost θ raises the cost of

[3] This reduced-form interpretation builds on work showing that cognitive uncertainty can affect choices and performance (Enke & Graeber, 2023), and on chess evidence that opponent gender can affect the quality of play in male-dominated competitive settings (Backus et al., 2023).

sustaining effort. Thus the same increase in stakes can generate different performance responses across opponent-gender environments.

Differentiating equation (10) with respect to I, holding M fixed, gives

$$\frac{\partial e^*(I,M)}{\partial I} = \frac{\alpha}{C_{ee}\left(e^*(I,M),\theta M\right)}. \tag{11}$$

Therefore, the local performance response to incentives is

$$\frac{\partial \mathbb{E}\left[p \mid I,M\right]}{\partial I} = \alpha \frac{\partial e^*(I,M)}{\partial I} = \frac{\alpha^2}{C_{ee}\left(e^*(I,M),\theta M\right)}. \tag{12}$$

Equation (12) is the core mechanism. The rating system observes realized performance, not latent ability. If the female-against-male environment attenuates the performance response to incentives, then high-stakes games in that environment generate lower realized performance even for the same latent ability.

Now compare a low-incentive state I_L to a high-incentive state $I_H > I_L$. Let

$$\Delta_0 \equiv \mathbb{E}\left[p \mid I_H, M=0\right] - \mathbb{E}\left[p \mid I_L, M=0\right] \tag{13}$$

denote the high-stakes response in female-opponent games, and let

$$\Delta_1 \equiv \mathbb{E}\left[p \mid I_H, M=1\right] - \mathbb{E}\left[p \mid I_L, M=1\right] \tag{14}$$

denote the high-stakes response in male-opponent games.

Define the relative high-stakes response as

$$\rho \equiv \frac{\Delta_1}{\Delta_0}. \tag{15}$$

If $0 \leq \rho \leq 1$ and $\Delta_0 > 0$, then $1-\rho$ is the share of the baseline high-stakes response that is attenuated in male-opponent games. If $\rho = 1$, the high-stakes response is the same across environments. If $\rho < 1$, the high-stakes response is smaller in female-against-male games.

3.3 A Quadratic Special Case

A special case helps fix intuition but is not required for identification. Suppose

$$C(e,\theta M) = \frac{1}{2}\left(1+\theta M\right)e^2. \tag{16}$$

Then

$$C_e(e,\theta M) = \left(1+\theta M\right)e, \qquad C_{ee}(e,\theta M) = 1+\theta M.$$

The interior effort choice is

$$e^*(I,M) = \frac{\alpha I}{1+\theta M}. \tag{17}$$

Expected performance is therefore

$$\mathbb{E}\left[p \mid I, M\right] = A_i + \frac{\alpha^2 I}{1 + \theta M}. \tag{18}$$

For female-opponent games, $M = 0$, so

$$\frac{\partial \mathbb{E}\left[p \mid I, M = 0\right]}{\partial I} = \alpha^2. \tag{19}$$

For male-opponent games, $M = 1$, so

$$\frac{\partial \mathbb{E}\left[p \mid I, M = 1\right]}{\partial I} = \frac{\alpha^2}{1 + \theta}. \tag{20}$$

Thus the relative response is

$$\rho = \frac{\Delta_1}{\Delta_0} = \frac{1}{1 + \theta}. \tag{21}$$

Equivalently,

$$\theta = \frac{1}{\rho} - 1. \tag{22}$$

This special case makes the interpretation transparent: a larger female-against-male cognitive cost θ flattens the performance response to incentives.[4]

3.4 Stylized Incorporation into Measured Ability

The framework above concerns realized performance. The next step is to note that rating systems do not observe latent ability A_i directly. They update measured ability from realized performance. A generic recursive measurement rule can be written as

$$R_{i,g+1} = R_{ig} + \kappa_g \left(p_{ig} - \widehat{p}_{ig}\right), \tag{23}$$

where R_{ig} is measured ability entering game g, $\widehat{p}_{ig}$ is the performance predicted by the rating system, and κ_g governs how quickly the system incorporates new performance information.

Taking expectations in equation (23) shows that any context-specific performance wedge is mechanically incorporated into the rating stock:

$$\mathbb{E}\left[\Delta R_{ig} \mid I, M\right] = \kappa_g \left(\mathbb{E}\left[p_{ig} \mid I, M\right] - \widehat{p}_{ig}\right). \tag{24}$$

Thus, if two players have the same latent ability but one realizes lower performance in a particular high-stakes environment, the rating system treats that lower realized performance as lower measured ability. The rule is objective and mechanical, but it updates from p_{ig} rather than from A_i.

Figure 3 illustrates how a local context-specific performance wedge can translate into a persistent

[4] However, the empirical exercise does not require estimating θ directly. The rating replay only requires the reduced-form performance wedge between the two environments.

distortion in measured ability. The simulation draws a population of players with fixed latent ability and feeds realized game outcomes through the USCF-style rating update rule, including the bonus-point logic. We use the quadratic cost case in equation (16) and set $\theta = 4$. Under this parameterization,

$$\rho = \frac{1}{1+\theta} = 0.2, \qquad 1 - \rho = 0.8.$$

Thus, the high-stakes performance response in female-against-male games is only 20 percent as large as the corresponding response in female-against-female games; equivalently, 80 percent of the baseline high-stakes response is attenuated. Panel A plots official Elo percentile against true latent-ability percentile for established players. The dashed 45-degree line is the no-distortion benchmark: players' official rating ranks coincide with their latent-ability ranks. The female calibration curve lies below this benchmark over much of the ability distribution, indicating that players with the same true latent rank receive lower official rating ranks when they are exposed to the context-specific performance cost. Panel B plots the distribution of the resulting rank penalty, defined as true latent-ability percentile minus official Elo percentile. Positive values therefore indicate that a player is ranked below their true latent-ability percentile. The simulation is not used for identification; it illustrates how the sufficient-statistic wedge can be mechanically incorporated into the rating stock and appear as a durable rank gap.

4 Data and Sample Construction

The analysis uses a game-level panel of rated USCF tournament games from 1994 to 2023. Each observation is a player-game. For each player-game, we observe the focal player's pre-game rating, the opponent's pre-game rating, the game result, event and section identifiers, round information, player gender, opponent gender, and demographic information when available. Game outcomes are coded as $S_{ig} \in \{0, 0.5, 1\}$, where wins receive one point, draws receive one-half point, and losses receive zero points.

The main empirical analysis focuses on established female focal players with at least 26 prior rated games and pre-game ratings between 800 and 2400. This restriction aligns the sample with the theoretical object of interest: whether female players' high-stakes performance response differs when the opponent is male rather than female. It also focuses the analysis on players for whom the standard rating-update rule is most relevant and avoids extreme rating ranges where institutional rules or competitive environments may differ.

Table 2 shows both the scale of the panel and the composition of the main identifying sample. The sample contains 978,135 female-focal player-games involving 8,037 female players, 71,585 events, and 114,226 event-sections. Female focal players face male opponents in 82.3 percent of games, reflecting the male-dominated composition of rated chess. Raw outcomes differ substantially by opponent gender: female focal players score 0.597 points per game against female opponents and 0.510 points per game against male opponents.

These raw differences also reflect rating and demographic composition. Female focal players

have a larger positive rating gap against female opponents than against male opponents, and male opponents are older on average. The empirical design therefore does not compare raw male- and female-opponent games directly. Instead, it estimates whether the near-threshold high-stakes response differs by opponent gender after conditioning on player fixed effects, rating-gap controls, and local tournament-state fixed effects.

Finally, bonus-threshold variables are constructed from the institutional rule described above. For each player-game, we reconstruct the player's pre-game distance to the bonus cutoff using the rating rule in force at the event date. These variables define the near-threshold treatment used in the reduced-form analysis and the affected games used in the counterfactual rating replay. Section 5 describes this construction in detail.

5 Empirical Design: Uncovering Context-Dependent Performance

In this section, we examine whether high-stakes incentives generate different performance responses across opponent-gender environments. The conceptual framework motivates this comparison directly: if female-against-male games impose an additional performance cost, then the same increase in incentives should produce a smaller score response in those games.

5.1 Empirical Variation

Our empirical variation comes from the USCF bonus-point rule. During an event, a player's cumulative performance can move close to the threshold at which an additional game score triggers a discrete bonus in the post-event rating update. We define the pre-game distance to the bonus cutoff in K units, where K is the player-specific rating-update parameter entering the USCF rule. The treated region is

$$\text{NearBelow}_{ieqr} = \mathbf{1}\left\{-1 \leq D_{ieqr} < 0\right\}, \tag{25}$$

where D_{ieqr} is the player's distance to the bonus cutoff before the game, measured in K units. Thus, $\text{NearBelow}_{ieqr} = 1$ identifies games in which the player is just below the bonus threshold and the current game is within reachable distance of triggering the bonus.

The choice of a one-K treated region is parsimonious and maps naturally to the rating rule: one additional game point has the potential to move the player across the event-level bonus threshold. Our preferred comparison sample uses games in a broader bracket around the cutoff, with the main specification comparing treated games in $[-1K, 0)$ to nearby comparison games outside the treated region. In this design, the omitted distance group consists of games in the comparison region, while the omitted opponent-gender group is female-opponent games. The coefficient of interest therefore compares how the high-stakes response changes when the female focal player faces a male opponent rather than a female opponent.

5.2 Empirical Specification and Coefficient Interpretation

For female focal players, we estimate game-level score regressions of the form

$$Y_{ieqr} = \beta_1 \text{NearBelow}_{ieqr} + \beta_2 \text{MaleOpp}_{ieqr} + \beta_3 (\text{NearBelow}_{ieqr} \times \text{MaleOpp}_{ieqr}) + f(\text{rating gap}_{ieqr}) + \alpha_i + \lambda_{e \times q \times r \times c} + \varepsilon_{ieqr}. \tag{26}$$

Here Y_{ieqr} is the game result score; i indexes the female focal player; e indexes the event; q indexes the section within event; r indexes the round; and c indexes the pre-round score cell. The variable MaleOpp_{ieqr} is an indicator for facing a male opponent. The function $f(\cdot)$ includes the rating gap and its square. Player fixed effects α_i absorb persistent differences in player ability and style. Event-section-round-score-cell fixed effects $\lambda_{e \times q \times r \times c}$ compare players in the same local tournament state.

The key comparison is a difference in high-stakes responsiveness across opponent-gender environments. In this specification, the estimated high-stakes response in female-opponent games is

$$\widehat{\Delta}_F = \widehat{\beta}_1, \tag{27}$$

while the estimated high-stakes response in male-opponent games is

$$\widehat{\Delta}_M = \widehat{\beta}_1 + \widehat{\beta}_3. \tag{28}$$

The interaction coefficient $\widehat{\beta}_3$ is therefore the central attenuation wedge. A negative $\widehat{\beta}_3$ means that high-stakes performance responsiveness is lower in male-opponent games. The corresponding score-unit wedge is $-\widehat{\beta}_3$, which we use later when translating the estimated performance wedge into rating implications.

To interpret magnitudes, we also report the estimated relative response

$$\widehat{\rho} = \frac{\widehat{\Delta}_M}{\widehat{\Delta}_F} = \frac{\widehat{\beta}_1 + \widehat{\beta}_3}{\widehat{\beta}_1}, \tag{29}$$

provided $\widehat{\beta}_1 > 0$. When $0 \leq \widehat{\rho} \leq 1$, the implied attenuation share is

$$1 - \widehat{\rho} = \frac{-\widehat{\beta}_3}{\widehat{\beta}_1}. \tag{30}$$

If $\widehat{\beta}_1 + \widehat{\beta}_3 < 0$, the net male-opponent response reverses sign, consistent with a cognitive-tax effect that dominates the ordinary incentive effect.

Because the data contain repeated games for the same players and common shocks within tournaments, our preferred inference clusters standard errors at the focal-player and event levels. This allows arbitrary correlation in unobserved performance shocks within player histories and within tournament environments.

5.3 Identification

The identifying assumption is that, absent the context-specific performance cost, the differential high-stakes response against male versus female opponents would evolve smoothly through the bonus-relevant region. In other words, conditional on player fixed effects, rating-gap controls, and event-section-round-score-cell fixed effects, female focal players just below the bonus threshold provide a valid local comparison to nearby female focal players outside the treated window.

The main threats to this assumption are sorting and composition around the cutoff. First, players close to the bonus threshold may differ in unobserved momentum, risk-taking, or match difficulty. Second, opponent gender may vary with tournament structure or score position, so treated games could mechanically involve different opponent pools. Third, because the bonus threshold is defined at the event level, distance to the cutoff is partly shaped by earlier performance and may correlate with recent shocks.

The empirical design addresses these concerns by making the comparison highly local within tournament state. Player fixed effects absorb persistent differences in ability and style. Event-section-round-score-cell fixed effects compare players in the same local tournament environment, round, and score position. Rating-gap controls absorb differences in expected score arising from relative strength. The remaining identifying variation is whether, within this local state, a female focal player is close enough to the bonus threshold for the current game to be high stakes and whether the opponent is male.

5.4 Diagnostics and Robustness

These restrictions do not mechanically rule out all forms of sorting around the cutoff, so we complement the main specification with design diagnostics and robustness checks. First, we document the support for the distance-to-cutoff design. Figure 4 plots the distribution of female-focal games by distance to the USCF bonus cutoff. Most usable observations are concentrated near the cutoff, and the treated region $[-1K, 0)$ accounts for a large share of the identifying variation. This supports the use of a parsimonious one-K treated window while preserving nearby comparison observations around the cutoff.

Second, we examine bin-level balance patterns around the bonus cutoff. Figure 5 plots mean result scores and the share of games against male opponents across distance-to-cutoff bins. The key diagnostic is whether the treated region is associated with a sharp change in opponent composition or baseline performance that could mechanically generate the estimated interaction. The bin-level patterns do not indicate a discrete break in opponent-gender composition at the treated region, supporting the interpretation that the main coefficient reflects differential high-stakes responsiveness rather than a discontinuous change in the opponent pool.

Third, we vary the treatment and comparison definitions. The preferred specification uses the treated region $[-1K, 0)$ and nearby comparison games around the cutoff. Figure 6 summarizes the alternative sample definitions used in robustness checks. We estimate below-only specifications that compare treated games only to games farther below the cutoff, donut specifications that drop

observations immediately above the cutoff, and more local bracket definitions such as $[-2K, 0.5K)$. These exercises test whether the estimated male-opponent attenuation wedge is driven by a particular choice of comparison window.

Fourth, we estimate a binned version of the design that replaces the single treatment indicator with distance-to-cutoff bins and interacts those bins with male-opponent status. This specification traces how the male-opponent wedge varies with the intensity of the bonus incentive. We normalize a pre-treatment bracket to zero and examine whether the interaction becomes most negative in the region immediately below the cutoff, where the bonus incentive is most salient.

6 Reduced-Form Evidence: Context-Dependent Performance Attenuation

We begin with reduced-form estimates of how female players' near-threshold performance response varies with opponent gender. The estimating equation compares games in which the focal female player is just below the bonus threshold to nearby comparison games, allowing the near-threshold response to differ between female- and male-opponent games. The key coefficient is the interaction between the near-threshold indicator and the male-opponent indicator.

Table 3 reports the main specification set. Across all specifications, the near-threshold response against female opponents is positive, while the interaction with male opponent is negative and statistically significant. Thus, the high-stakes response that appears against female opponents is substantially attenuated when the opponent is male.

The preferred specification is column (6), which includes player fixed effects and score-cell fixed effects. In this specification, the estimated near-threshold response against female opponents is 0.0116, while the interaction with male opponent is -0.0098. The implied near-threshold response against male opponents is therefore only

$$0.0116 - 0.0098 = 0.0018.$$

This pattern is the central reduced-form result: female players increase performance near the bonus threshold, but this response is almost fully muted in male-opponent games.

The interaction coefficient also gives the sufficient-statistic input used in the replay exercise. Let β_1 denote the near-threshold response against female opponents, and let β_3 denote the additional interaction effect in male-opponent games. The response against male opponents is $\beta_1 + \beta_3$. We define

$$\widehat{\rho} = \frac{\widehat{\beta}_1 + \widehat{\beta}_3}{\widehat{\beta}_1},$$

where $\widehat{\rho}$ is the share of the near-threshold response that remains in male-opponent games. The implied attenuation is

$$1 - \widehat{\rho} = \frac{-\widehat{\beta}_3}{\widehat{\beta}_1}.$$

In the preferred specification,

$$\widehat{\rho} = \frac{0.0116 - 0.0098}{0.0116} \approx 0.155, \qquad 1 - \widehat{\rho} \approx 0.845.$$

Thus, roughly 84.5% of the near-threshold response is attenuated in male-opponent games. Equivalently, the score-unit wedge used in the replay is

$$\widehat{\delta} = -\widehat{\beta}_3 = 0.0098.$$

This object has a direct interpretation: it is the estimated score loss, in result-score units, associated with being in the bonus-relevant region against a male opponent, relative to the corresponding near-threshold response against a female opponent.

Result 1 (Context-dependent performance attenuation)**.** *The main estimates show economically meaningful context-dependent attenuation in female focal players' high-stakes performance response. In the preferred specification, the high-stakes response is* 0.0116 *score units against female opponents, but the male-opponent interaction is* -0.0098*; this implies that roughly 84 percent of the baseline high-stakes response is attenuated in female-against-male games.*

Table 4 shows that this result is stable across alternative definitions of the comparison group. The estimated interaction remains negative and statistically significant in all specifications, with magnitudes ranging from -0.0087 to -0.0104. The implied attenuation is also similar across specifications. Most estimates imply that male-opponent games attenuate roughly 75.4% to 84.5% of the near-threshold response; the donut specification implies nearly complete attenuation.

Figure 7 provides the corresponding visual and nonparametric check across alternative fixed-effect specifications. Instead of imposing a single near-threshold indicator, the figure estimates male-opponent interaction coefficients separately by distance-to-cutoff bins, normalizing the omitted bin $[-2K, -1K)$ to zero. The pattern is concentrated where the design predicts: the male-opponent wedge becomes most negative just below the bonus threshold, especially in the $[-1K, 0)$ region. By contrast, the estimates are closer to zero farther from the cutoff and above the threshold. This supports the interpretation that the main interaction is tied to the bonus-relevant state rather than reflecting a smooth gender-opponent difference throughout the rating distribution.

7 Counterfactual Ratings: From *Flow* to *Stock* Distortion

The reduced-form results identify a local *flow* distortion in realized game scores: female players exhibit an attenuated high-stakes response when the opponent is male. The natural next question is whether this local score-unit wedge is merely transitory, or whether the rating system mechanically incorporates it into the stock of measured ability. We answer this question by replaying the observed rating rule under corrected score histories.

7.1 Counterfactual replay and rating incorporation gaps

Let the estimated score-unit attenuation be

$$\widehat{\delta} \equiv -\widehat{\beta}_3 > 0,$$

where $\widehat{\beta}_3$ is the estimated near-threshold-by-male-opponent interaction from the preferred reduced-form specification. Define affected games as female-focal games against male opponents in the bonus-relevant window:

$$\text{Affected}_{ig} = \mathbb{1}\{\text{Male}_{ig} = 1\} \times \mathbb{1}\{-K_{ie} \leq D_{ier} < 0\}. \tag{31}$$

For these games, the counterfactual score adds back the estimated attenuation wedge:

$$S_{ig}^{cf} = \min\left\{1, \max\left\{0, S_{ig}^{obs} + \widehat{\delta}\text{Affected}_{ig}\right\}\right\}. \tag{32}$$

For non-affected games, $S_{ig}^{cf} = S_{ig}^{obs}$. The clipping in equation (32) preserves the feasible per-game score interval $[0, 1]$, where observed chess outcomes take values 0, 0.5, or 1.

The replay then asks how the observed rating rule would have processed this corrected score history. In the baseline implementation, expected scores are held fixed at their observed values. This choice isolates the mechanical incorporation channel: the exercise does not ask how pairings, effort choices, or future outcomes would have changed under a different rating path. Instead, it asks how much of the estimated flow wedge would be converted into ratings by the rule actually used in the data.

Let

$$S_{ie}^{cf} = \sum_{g \in e} S_{ig}^{cf}, \qquad S_{ie}^{obs} = \sum_{g \in e} S_{ig}^{obs}, \qquad E_{ie}^{obs} = \sum_{g \in e} E_{ig}^{obs}.$$

The counterfactual event-level update is

$$\Delta R_{ie}^{cf} = K_{ie}(S_{ie}^{cf} - E_{ie}^{obs}) + \max\left\{0, K_{ie}(S_{ie}^{cf} - E_{ie}^{obs}) - B_t\sqrt{m'_{ie}}\right\}. \tag{33}$$

The corresponding observed mechanical update is

$$\Delta R_{ie}^{obs} = K_{ie}(S_{ie}^{obs} - E_{ie}^{obs}) + \max\left\{0, K_{ie}(S_{ie}^{obs} - E_{ie}^{obs}) - B_t\sqrt{m'_{ie}}\right\}. \tag{34}$$

The event-level incorporation gap is therefore

$$\text{Gap}_{ie} = \Delta R_{ie}^{cf} - \Delta R_{ie}^{obs}. \tag{35}$$

Equivalently, this gap can be written as the sum of the ordinary K-factor update generated by the

corrected score and any additional change induced by the bonus rule.[5]

Cumulative incorporation before event e is

$$\text{Gap}_{ie}^{cum} = \sum_{e'<e} \text{Gap}_{ie'}, \qquad R_{ie}^{cf} = R_{ie}^{obs} + \text{Gap}_{ie}^{cum}. \tag{36}$$

The player-level incorporation gap is the difference between the counterfactual and observed rating paths accumulated over the player's observed history.

This replay is intentionally mechanical and partial equilibrium. It does not simulate new pairings, reassign wins and losses, model entry or attrition, or estimate a structural effort-cost parameter. The object is narrower: holding fixed the estimated score wedge, how much would measured ratings change if the observed rating rule were applied to corrected score histories? The resulting estimates should therefore be interpreted as rating-rule incorporation effects, not as a full counterfactual tournament equilibrium.

Table 6 reports the player-level incorporation gaps. The average counterfactual rating gain is 15.56 Elo points, with a median of 7.88 points and a 90th percentile of 41.48 points. The exposure rows show that the *stock* distortion accumulates through repeated exposure. Players with only one to five affected games have an average gap of 1.02 Elo points. Players with more than 20 affected games have an average gap of 25.15 Elo points and a 90th percentile gap of 54.71 Elo points. Thus, the replay does not imply a uniform shift in measured ratings. Instead, the rating consequences are concentrated among players whose competitive histories repeatedly place them in the affected state.

Figure 8 plots the full player-level distribution of replay-implied rating gaps. The distribution is strongly right-skewed. Most players have positive but relatively modest incorporation gaps, while a smaller group accumulates substantially larger rating losses from repeated exposure. The dashed vertical line marks the sample mean. The fact that the mean lies to the right of the bulk of the distribution reflects the influence of this upper tail.

The magnitude of these gaps is meaningful because ratings are the state variable through which past performance is summarized and carried forward. The average gap is moderate, but it is not a one-game score effect; it is a persistent difference in measured ability. Moreover, the distribution is highly right-skewed. For players in the upper tail of exposure, the replay-implied gaps are large enough to affect future expected scores, tournament classifications, and the interpretation of subsequent performance. The key implication is that once a context-dependent *flow* distortion enters the rating update, it can become part of the institutional record of ability.

Result 2 (Rating incorporation)**.** *Applying the observed rating rule to corrected score histories implies an average player-level rating gap of 15.56 rating points, with a median gap of 7.88 points*

[5] In the current implementation, most of the mean player-level incorporation gap operates through ordinary K-factor updating rather than the bonus rule: 13.71 Elo points, or 88.1%, come from the base update, while 1.86 Elo points, or 11.9%, come from bonus amplification. We treat this as a mechanical accounting result rather than the main economic decomposition. The bonus threshold is central for identifying the local performance wedge; the replay then shows how the broader rating rule incorporates that wedge into measured ability.

and a 90th percentile gap of 41.48 points. The distortion is concentrated among repeatedly exposed players: players with more than 20 affected games have an average replay-implied gap of 25.15 points and a 90th percentile gap of 54.71 points.

7.2 Decomposing current flow losses from rating-stock losses

The replay also allows us to separate the contemporaneous flow wedge from the accumulated *stock* distortion. This distinction matters because ratings are not only controls for ability in later games. They are themselves cumulative outcomes of previous measured performance. If past performance was distorted in a context-dependent way, then the observed rating already embeds part of that history before the next game begins.

To measure this embedded component, we compare the expected score implied by the observed rating to the expected score implied by the replay-corrected rating. Before game g, define

$$\text{Rating-absorbed loss}_{ig} = E\left(R_{ie}^{obs} + \text{Gap}_{ie}^{cum} - R_{ig}^{opp}\right) - E\left(R_{ie}^{obs} - R_{ig}^{opp}\right). \tag{37}$$

This object measures the expected-score gain that would arise if the player's rating stock already reflected the corrected historical score path. Equivalently, it is the expected-score loss embedded in the observed rating stock. It is a stock component: it comes from past affected games that have already been processed through the rating rule.

The direct current-game loss is

$$\text{Direct current loss}_{ig} = \widehat{\delta}\text{Affected}_{ig}. \tag{38}$$

This object is a flow component: it is nonzero only when the current game itself is in the affected state. The total replay-implied loss is therefore

$$\text{Total replay-implied loss}_{ig} = \text{Rating-absorbed loss}_{ig} + \text{Direct current loss}_{ig}. \tag{39}$$

Importantly, this decomposition differs from a standard rating-controlled male-opponent regression. Such a regression asks whether, conditional on the observed rating gap and other controls, current outcomes differ when the opponent is male. That is a useful descriptive object, but it is not the object of the replay. Observed ratings are predetermined relative to the current game, yet they are themselves cumulative outcomes of earlier measured performance. If earlier affected-game scores were attenuated and then incorporated into ratings, conditioning on observed ratings can absorb part of the mechanism of interest. The replay instead starts from the locally identified score wedge and uses the institutional updating rule to separate the current flow loss from the portion already embedded in the rating stock.[6]

[6]For example, a rating-gap-controlled regression of the form $Y_{ig} = \delta\text{Male}_{ig} + f(R_{ig}^{obs} - R_{ig}^{opp}) + \alpha_i + \lambda_{e,r,q,s} + \varepsilon_{ig}$ estimates a residual male-opponent penalty conditional on the observed rating gap. By contrast, the replay asks how the estimated flow distortion would have changed the rating path itself, and therefore how much of the replay-implied loss appears as a current-game flow versus an accumulated rating-stock component.

Table 7 reports the resulting decomposition. Among all male-opponent games, 87.5% of the replay-implied loss is already embedded in ratings before the current game begins. This broad sample includes many male-opponent games that are not themselves in the affected window, so the direct current-game component is naturally smaller. Among affected games only, where the contemporaneous flow wedge is active by construction, the absorbed share remains 79.7%. Thus, even in games where the current *flow* distortion is present, most of the replay-implied loss comes from the rating stock rather than the current game alone.

Result 3 (Flow-to-stock decomposition)**.** *Most of the replay-implied loss operates through the rating stock rather than the current game alone. Among all male-opponent games, 87.5% of the replay-implied loss is already embedded in ratings before the current game begins. Among affected games, where the contemporaneous flow wedge is active by construction, the absorbed stock component remains 79.7%.*

7.3 Summary

This decomposition is the main bridge from *flow* distortion to *stock* distortion. The reduced-form estimates show that performance is attenuated in a specific high-stakes context. The replay shows that this attenuation is not fully reset after the game ends. Because the rating rule uses realized scores to update measured ability, past context-dependent performance losses become embedded in the rating stock. Subsequent games then begin from a rating that already reflects these accumulated losses. In this sense, observed ratings are not merely predetermined controls for ability; they are institutional summaries of the environments in which prior performance was measured.

Taken together, the replay and decomposition should be interpreted as a partial-equilibrium measure of rating-rule incorporation. The exercise holds fixed the observed sequence of opponents, expected scores, entry, attrition, and effort responses. It therefore does not estimate the full equilibrium consequences of removing the performance wedge. Instead, it isolates the institutional object most directly tied to the paper's central claim: conditional on the observed tournament history and the estimated flow distortion, how much of that distortion is mechanically transformed into the stock of measured ability?

8 Design Implications: the Equity–Efficiency Frontier

The previous section shows that context-specific performance attenuation can be incorporated into ratings through the mechanical updating process. We now use the same framework to study a design question: holding fixed the estimated performance wedge, how should an institution choose the speed at which observed performance is capitalized into measured ability?

This exercise is not meant to recover a single optimal rating rule. Rather, it uses the rating system as a laboratory for studying a broader policy trade-off. A faster updating rule learns more quickly from new performance information, but it also incorporates distorted performance signals

more aggressively. A slower updating rule provides partial insurance against transitory or context-specific shocks, but it delays convergence between measured ratings and latent ability. Thus, the update parameter governs both efficiency and equity.

8.1 Rating updates as institutional design

In each game, ratings are updated according to the difference between the realized score and the expected score:

$$R_{i,g+1} = R_{ig} + K_{ig}\,(S_{ig} - E_{ig}) + B_{ie(g)}, \tag{40}$$

where $S_{ig} \in \{0, 0.5, 1\}$ is the realized game score, E_{ig} is the expected score implied by the current rating gap, K_{ig} is the game-level update weight, and $B_{ie(g)}$ denotes any event-level bonus adjustment.

The status quo USCF rule makes K_{ig} larger for players with fewer prior rated games and smaller for more established players. In the counterfactuals below, we treat the status quo rule as the benchmark and compare it to a menu of alternative schedules. The key design question is whether a schedule can reduce the incorporation of the attenuation wedge without greatly slowing convergence to latent ability. [7]

The affected-game discount schedules are useful for mechanism and benchmarking purposes, but they are not part of the main frontier. They require the institution to identify the context in which performance is attenuated. The main frontier instead focuses on rules that can be implemented using only ratings and game histories.

8.2 Measuring bias incorporation and learning speed

For each counterfactual schedule $K^{cf} \in \mathcal{K}$, we replay the same game history twice. The first replay uses the observed scores. The second replay adds back the estimated attenuation wedge in affected games:

$$S_{ig}^{+} = S_{ig} + \widehat{\Delta} \cdot \mathbf{1}\{(i,g) \in \mathcal{A}\}, \qquad \widehat{\Delta} = -\widehat{\beta}_3, \tag{41}$$

where $\mathcal{A}$ denotes affected games and $\widehat{\beta}_3$ is the estimated near-threshold-by-male-opponent interaction from the preferred reduced-form specification. In implementation, the correction is bounded so that scores remain in the feasible score interval.

We summarize each counterfactual schedule using two objects. The first is

$$\text{IncorporationBias}(K^{cf}),$$

defined as the mean player-level rating gain from replaying the rating rule after adding back the attenuation wedge. Equivalently, it measures how much of the observed score attenuation would be

[7] The event-level bonus rule reinforces this logic. Because the bonus is triggered by cumulative rating gains within an event, increasing K affects not only the direct game-level update $K(S - E)$, but also the probability and size of bonus adjustments. In the counterfactual replay, we therefore decompose the total rating penalty into a base-update component and a bonus component.

incorporated into durable ratings under schedule K^{cf}. Larger values indicate stronger pass-through from context-dependent performance attenuation into measured ability.

The second object is

$$\text{LearningSpeed}(K^{cf};\eta),$$

defined as the median number of games required for ratings to converge near latent ability. Because this part of the exercise is simulated, we observe each player's true latent ability A_i. We compare the player's rating percentile to the player's latent-ability percentile and measure how many games are required for the two to remain within η percentage points of each other. Our preferred benchmark uses $\eta = 5$ percentage points and requires convergence to be sustained over a ten-game window.

The equity–efficiency frontier is then defined over the pairs

$$\Big(\text{LearningSpeed}(K^{cf};\eta), \text{IncorporationBias}(K^{cf})\Big) \quad \text{for} \quad K^{cf} \in \mathcal{K}.$$

Both dimensions are costs. Lower learning speed means fewer games are required to learn latent ability, while lower incorporation bias means less of the attenuation wedge is capitalized into ratings. A schedule is dominated if another schedule is both faster and incorporates less bias. The frontier consists of the non-dominated schedules.

Figure 9 shows the basic trade-off. The observed USCF rule is relatively fast, but it also incorporates a large share of the attenuation wedge. In the current simulation, the observed rule implies a mean attenuation rating penalty of about 15 rating points, while lower-update schedules such as flat $K = 8$ or flat $K = 12$ reduce the corresponding penalty to roughly 1–2 rating points. The cost is slower convergence: low-update schedules require more games to bring measured ratings close to latent ability.

Scaled versions of the current rule illustrate the same logic. Reducing the current rule to 50 percent preserves much of the speed of the benchmark in the simulation, but substantially reduces the rating penalty. More aggressive reductions lower incorporation further, but increasingly slow learning. The choice of K therefore cannot be evaluated only by asking how quickly ratings respond to new information. The same responsiveness determines how quickly distorted information becomes part of a player's measured ability stock.

Figure 10 traces the lower-left frontier. Some schedules are dominated: they require more games to learn latent ability and incorporate more attenuation bias than available alternatives. Other schedules lie closer to the frontier, improving one objective while sacrificing relatively little of the other. The status quo is therefore not neutral. It represents one institutional choice about how quickly observed performance is transformed into durable ranking capital.

Result 4 (Rating-system design)**.** *Holding fixed the estimated score wedge, faster update schedules generally learn latent ability more quickly, but also incorporate more of the context-specific attenuation wedge into ratings. Lower or delayed update schedules reduce bias pass-through, but at the cost of slower convergence.*

This exercise has two implications. First, rating systems face an equity–efficiency trade-off even

when the rating formula is mechanically objective. If the input signal is distorted, faster learning can also mean faster capitalization of unequal performance costs. Second, design improvements need not require changing the scoring rule itself. Institutions can change how quickly scores are incorporated, whether update rates are capped, and whether early-career ratings should be protected from high-variance environments.

The exercise is partial equilibrium. It holds fixed the estimated performance wedge, the pairing environment, and players' behavioral responses. It therefore does not ask how players would change effort, entry, or tournament selection under alternative rules. The purpose is narrower: to show that once observed performance contains context-dependent attenuation, the update rule itself becomes a policy instrument. In this sense, K determines both how fast ratings learn and how fast distortions become ratings.

9 Conclusion

This paper studies how measured talent can differ from latent ability when performance is evaluated through an objective but context-dependent measurement system. The empirical setting provides a clean laboratory because chess outcomes are objectively scored, repeated over time, and transformed into ratings through an explicit institutional rule. We show that female players exhibit an attenuated high-stakes response when facing male opponents near bonus-relevant thresholds. Under the sufficient-statistic interpretation, $\widehat{\rho}$ summarizes relative high-stakes responsiveness under general convex performance costs, while $-\widehat{\beta}_3$ provides the score-unit wedge used in the rating replay.

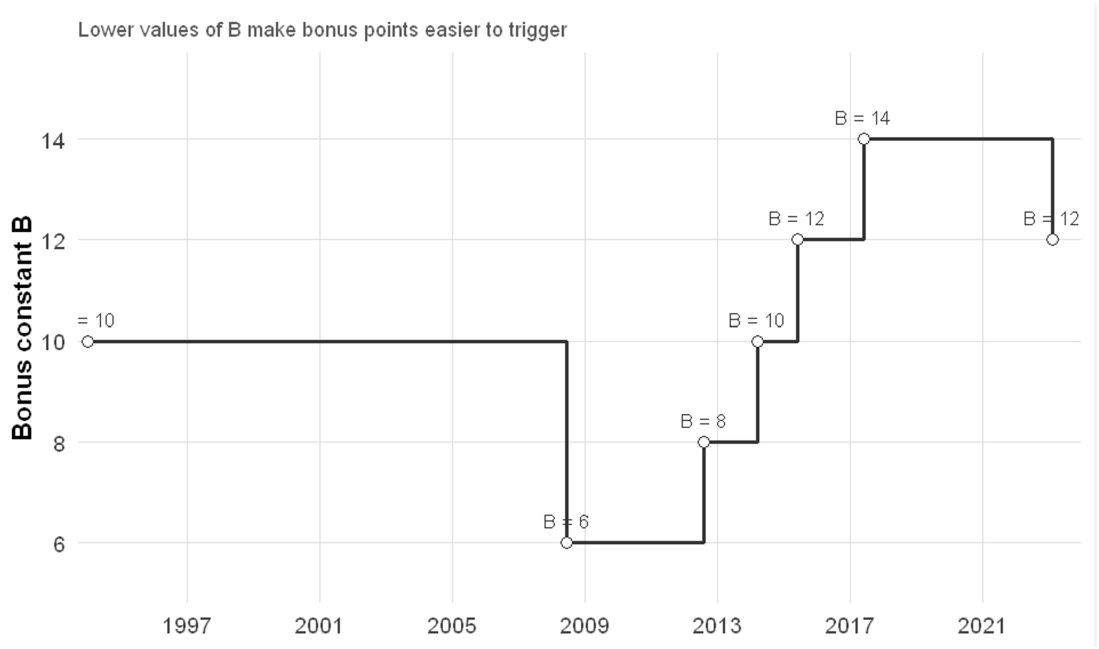
The counterfactual replay shows that this local flow wedge is not merely transitory. Because ratings are updated from realized scores, context-dependent performance losses can be mechanically incorporated into the stock of measured ability. The replay-implied rating gaps are moderate on average but highly right-skewed, with larger accumulated penalties among players repeatedly exposed to the affected state. The stock-flow decomposition further shows that much of the replay-implied loss is already embedded in ratings before the current game begins. Thus, observed ratings are not only controls for ability; they are institutional summaries of the environments in which prior performance was measured.

The final counterfactual exercises use this framework to study rating-system design. Holding fixed the estimated score wedge, faster update schedules learn latent ability more quickly but also incorporate distorted signals more aggressively. Slower or delayed update schedules reduce bias pass-through but slow convergence. The exercise does not identify a universal optimal update rule. Instead, it highlights a broader equity–efficiency frontier: measurement systems must choose how quickly to capitalize observed performance when observed performance may itself reflect unequal or context-dependent costs.

More broadly, the paper contributes to the economics of talent measurement by showing that objectivity at the outcome level does not guarantee neutrality in measured talent. Even when

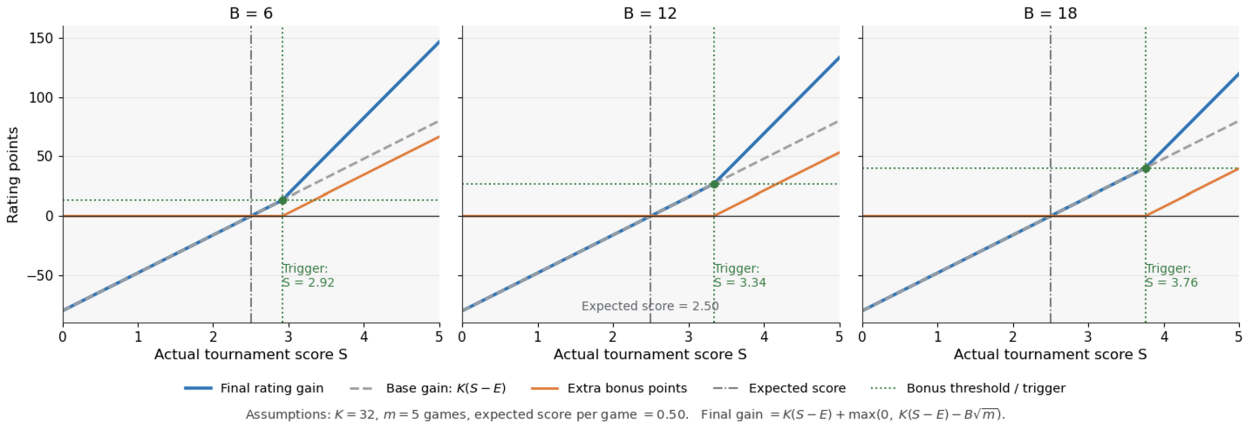
outcomes are precisely scored and updating rules are transparent, measured ability can reflect both latent skill and the historical conditions under which that skill was evaluated. This logic extends beyond chess to other settings in which high-stakes performance is converted into durable rankings, credentials, or evaluations. Designing better measurement systems therefore requires attention not only to how accurately institutions learn from performance, but also to what kinds of performance signals they choose to embed into the record of talent.

Figure 1: USCF bonus constant over time



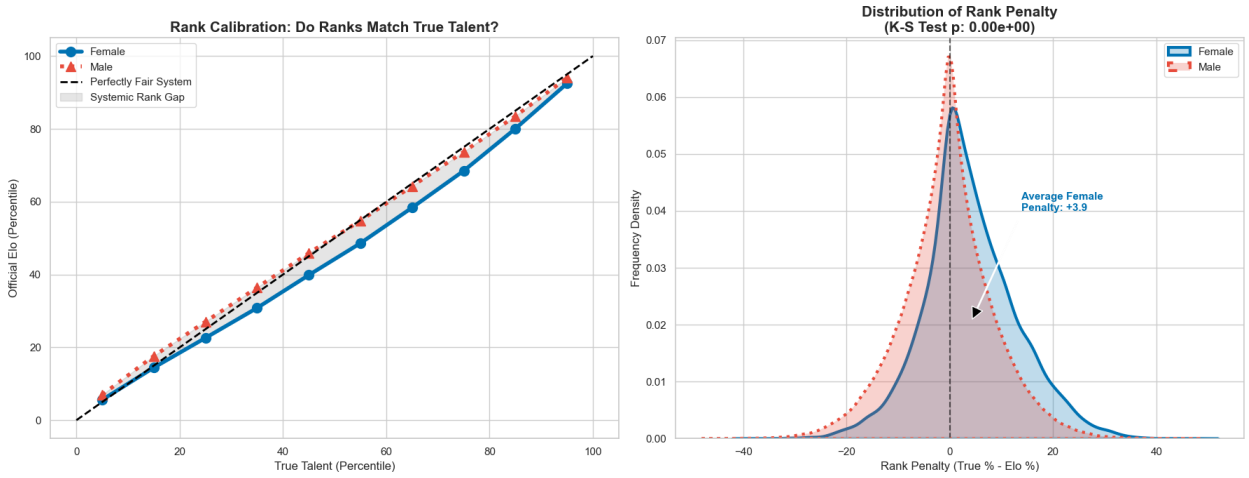
Notes: The figure plots the USCF bonus constant B_t over time. Lower values of B_t make bonus points easier to trigger, while higher values make bonus points harder to earn. Under the established-player update rule, bonus points are awarded when $K_{ie}(S_{ie} - E_{ie}) > B_t\sqrt{m'_{ie}}$, where K_{ie} is the event-specific K-factor, $S_{ie} - E_{ie}$ is event-level over-performance relative to expected score, and $m'_{ie} = \max\{m_{ie}, 4\}$. The schedule shown here highlights that the institutional threshold varied over time during the sample period.

Figure 2: Mechanics of the USCF Bonus-Point Rule



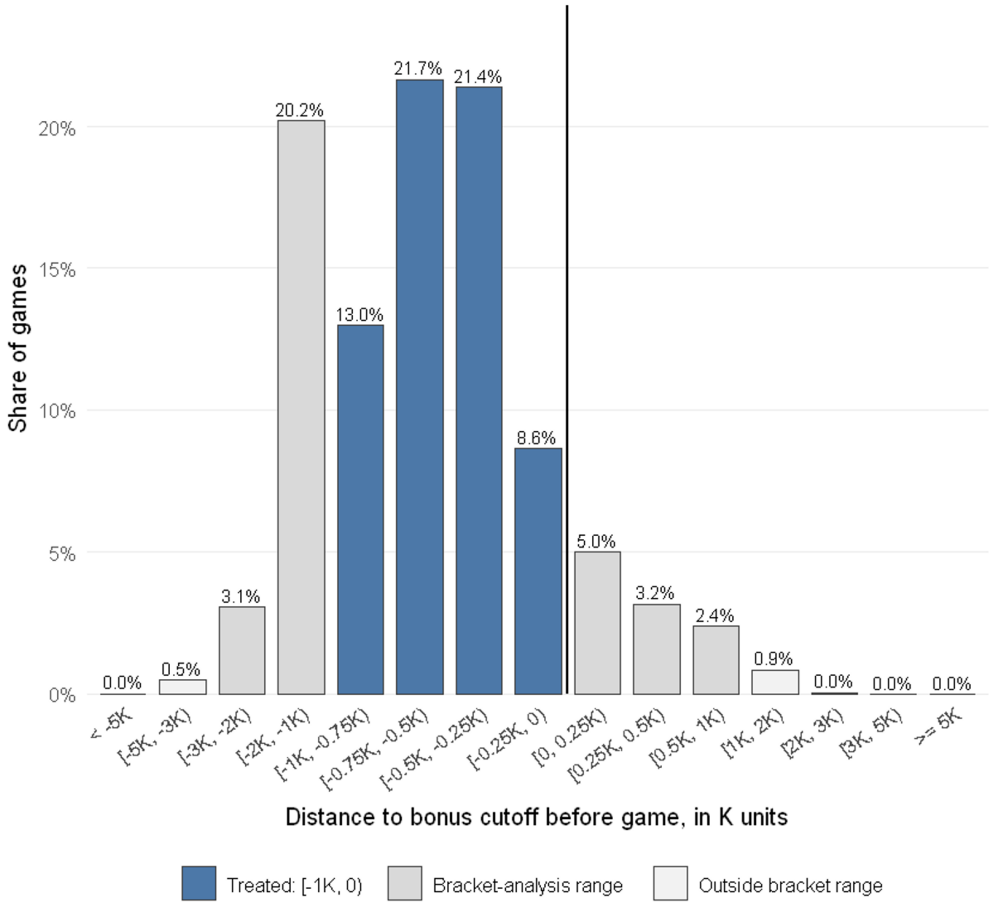
Notes: The figure illustrates how the USCF bonus-point rule converts tournament performance into additional rating points. Each panel fixes the number of games at $m = 5$, the update parameter at $K = 32$, and the expected score at $E = 2.5$, while varying the bonus threshold parameter B. The dashed gray line plots the base rating gain, $K(S - E)$, as a function of the player's tournament score S. The orange line plots the extra bonus points, which are zero until the base gain exceeds the bonus threshold, $B\sqrt{m}$, and positive thereafter. The blue line plots the final rating gain, $K(S-E)+\max 0, K(S-E) - B\sqrt{m}$. Because game scores are discrete, the realized tournament score S is observed only in half-point increments, $S \in 0, 0.5, \ldots, m$. The plotted lines linearly interpolate between these feasible score values for exposition. The vertical dash-dotted line marks the expected score, and the green dotted lines mark the continuous bonus threshold and the corresponding trigger score; in implementation, the realized trigger occurs at the first feasible half-point score above the threshold.

Figure 3: Stylized Rating Distortion Under Context-Specific Performance Costs



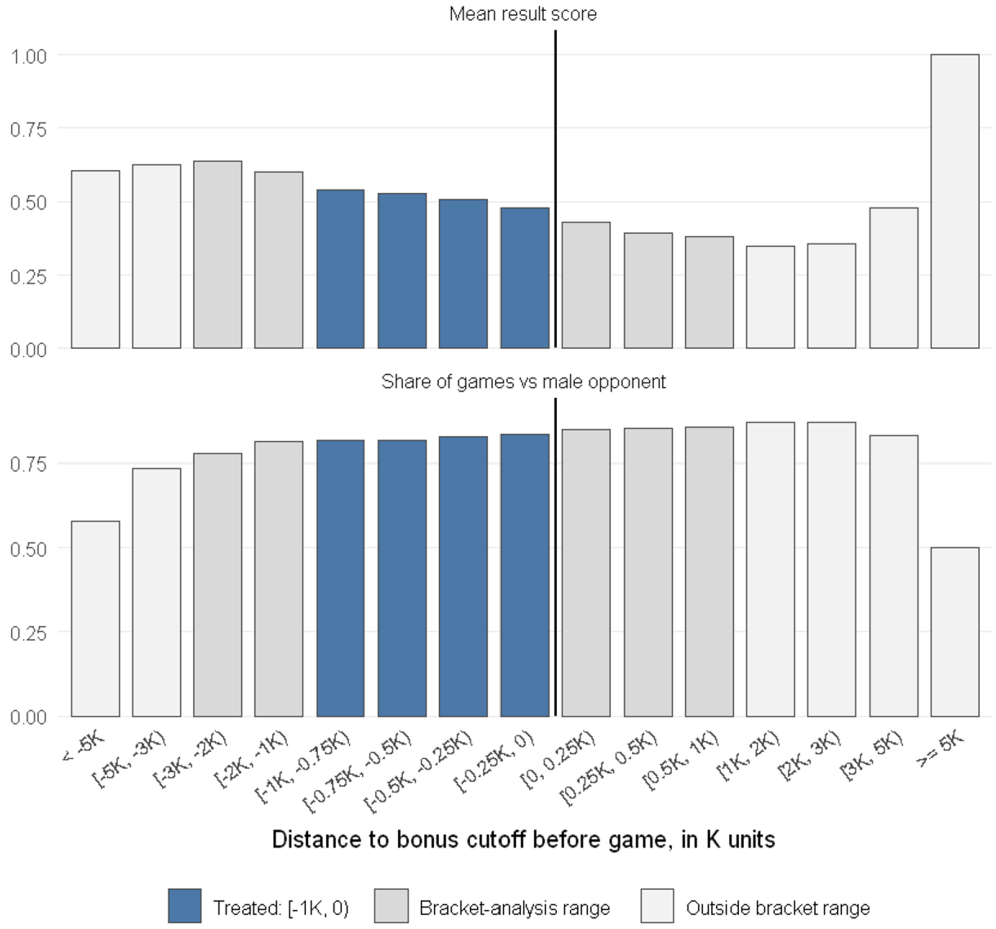
Notes: The figure reports a stylized simulation with fixed latent ability, quadratic effort costs, and the USCF-style rating update rule. The simulation sets $\theta = 4$, which implies $\rho = 1/(1+\theta) = 0.2$ and hence an 80 percent attenuation of the high-stakes performance response in female-against-male games. Panel A plots official Elo percentile against true latent-ability percentile for established players; the dashed line is the no-distortion benchmark. Panel B plots the distribution of the rank penalty, defined as true latent-ability percentile minus official Elo percentile. Positive values indicate that the player's official rating rank is below the player's true latent-ability rank.

Figure 4: Sample Support Around the Bonus Cutoff



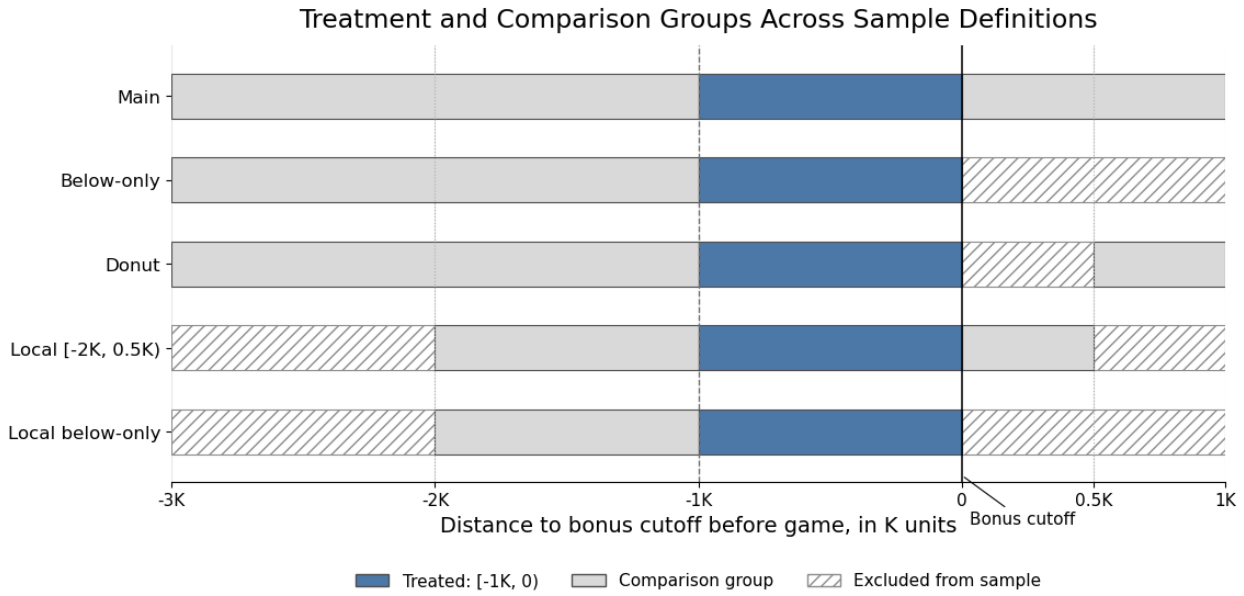
Notes: The figure reports the distribution of female-focal games by distance to the USCF bonus cutoff before the game, measured in K units. Bars show the share of nonmissing games in each distance bin. The blue bars denote the treated region, defined as games in which the female focal player is within one K unit below the bonus cutoff, $[-1K, 0)$. Dark gray bars denote the bracket-analysis range used for nearby comparison games, and light gray bars denote bins outside the bracket range. The vertical line marks the bonus cutoff.

Figure 5: Balance Diagnostics Around the Bonus Cutoff



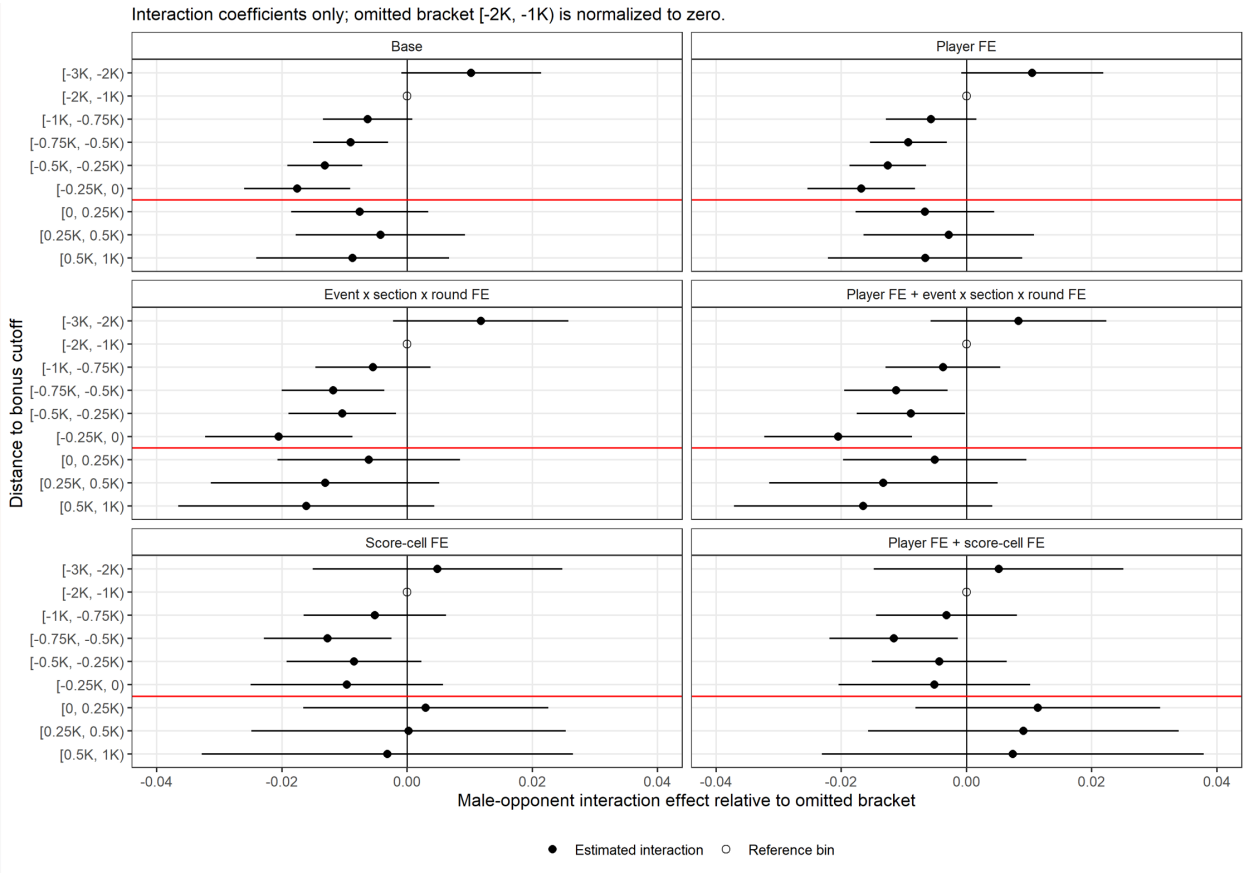
Notes: The figure reports bin-level diagnostics by distance to the USCF bonus cutoff before the game, measured in K units. The top panel plots the mean game result score within each distance bin, and the bottom panel plots the share of games played against male opponents. The blue bars denote the treated region, defined as games in which the female focal player is within one K unit below the bonus cutoff, $[-1K, 0)$. Dark gray bars denote the bracket-analysis range used for nearby comparison games, and light gray bars denote bins outside the bracket range. The vertical line marks the bonus cutoff.

Figure 6: Treatment and Comparison Windows Around the Bonus Cutoff



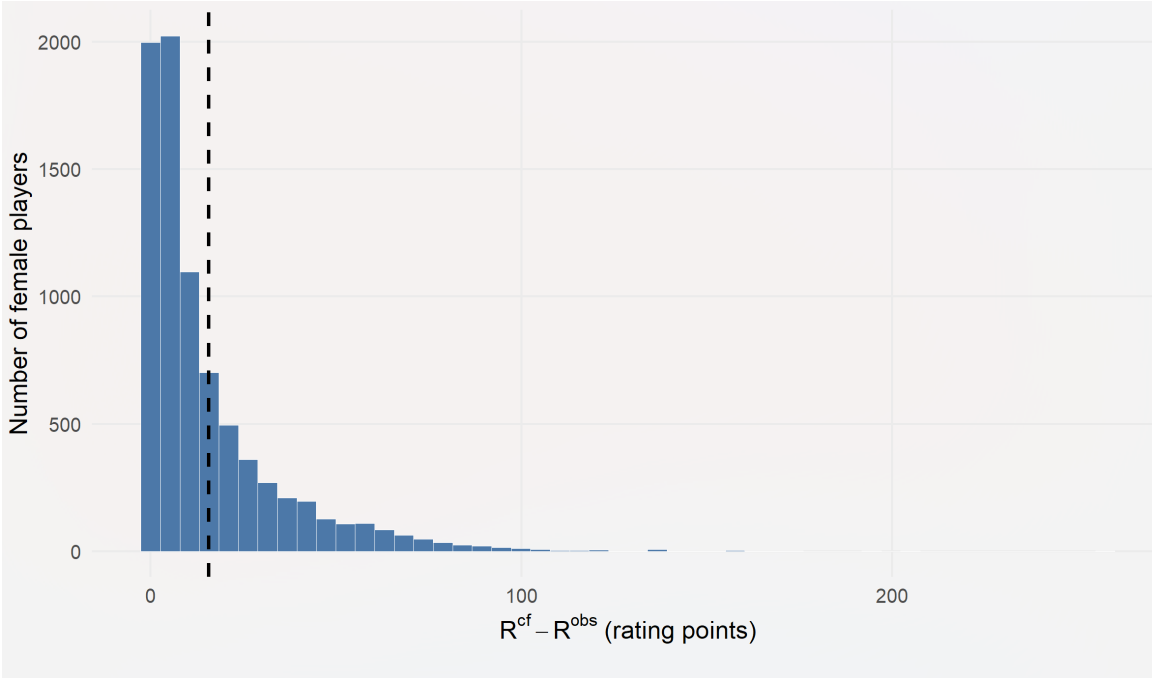
Notes: The figure illustrates the treatment and comparison windows used across the main specification and robustness samples. The horizontal axis reports distance to the USCF bonus cutoff before the game, measured in K units. The treated region is fixed across specifications and consists of games in which the female focal player is within one K unit below the bonus cutoff, $[-1K, 0)$. Gray regions denote comparison observations included in each sample definition, while hatched regions denote observations excluded from that specification. The solid vertical line marks the bonus cutoff at zero, and the dashed vertical line marks the lower boundary of the treated region at $-1K$. The main specification uses a broad bracket around the cutoff; the below-only, donut, and local specifications vary the comparison window to test whether the estimated male-opponent attenuation wedge is sensitive to the choice of nearby control games.

Figure 7: Male-Opponent Attenuation Wedge by Distance to the Bonus Cutoff



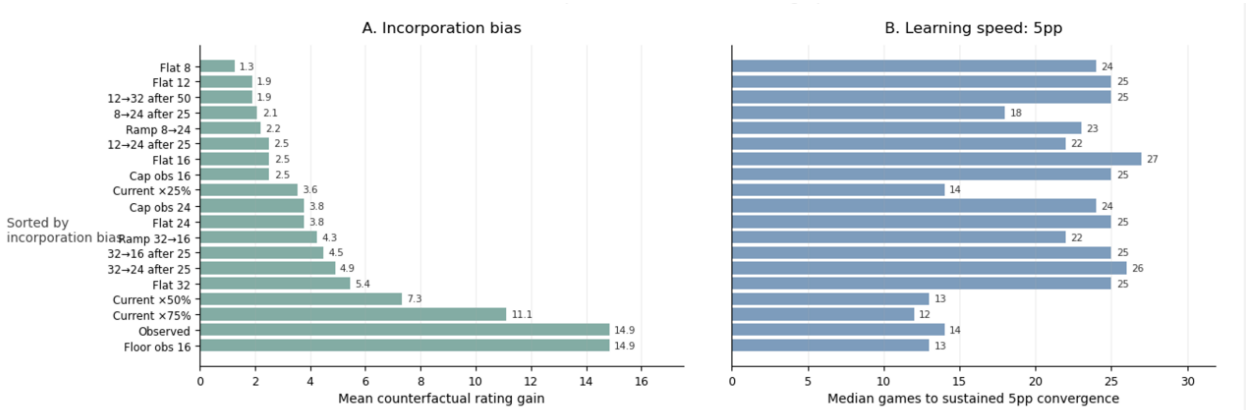
Notes: The figure reports binned estimates of the male-opponent interaction effect for female focal players across alternative specifications. Each point is the estimated coefficient on the interaction between a distance-to-bonus-cutoff bin and an indicator for facing a male opponent; horizontal lines denote 95 percent confidence intervals. The omitted reference bin is $[-2K, -1K)$, so all coefficients are interpreted relative to that bracket. The red horizontal line marks the bonus cutoff, separating games below the threshold from games above it. Panels differ by the included fixed effects: the baseline specification, player fixed effects, event-by-section-by-round fixed effects, player fixed effects combined with event-by-section-by-round fixed effects, score-cell fixed effects, and player fixed effects combined with score-cell fixed effects. Negative coefficients in bins just below the cutoff indicate that the high-stakes performance response is weaker in male-opponent games precisely where bonus incentives are most salient.

Figure 8: Distribution of replay-implied rating incorporation gaps



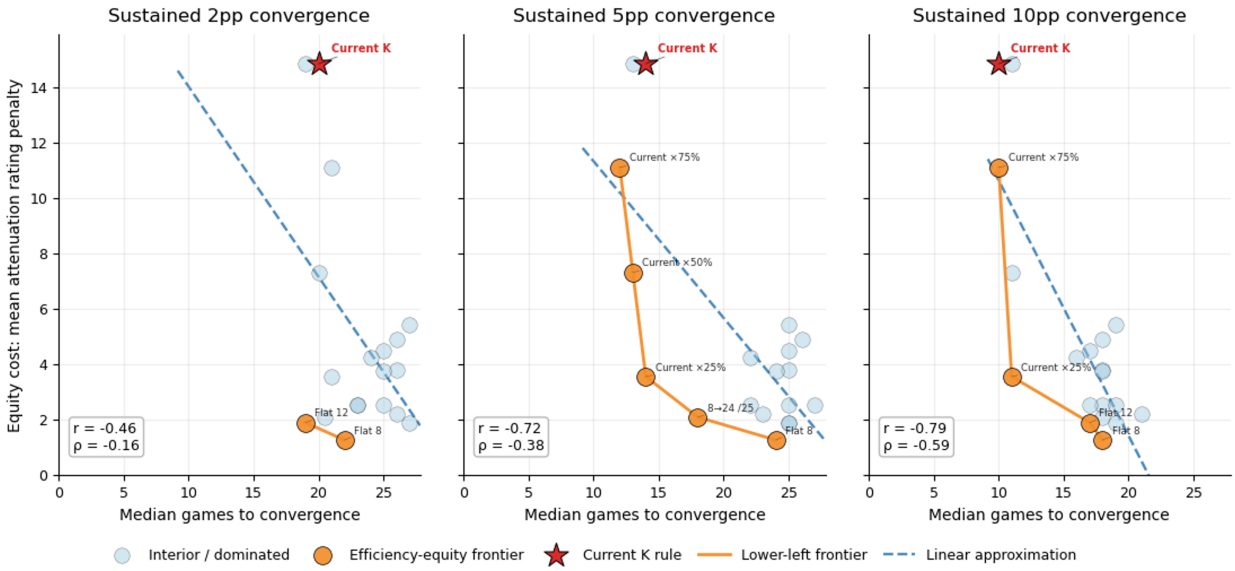
Notes: The figure plots the player-level distribution of $R^{cf} - R^{obs}$ from the baseline replay. The counterfactual rating R^{cf} is constructed by adding back the estimated near-threshold-by-male-opponent score wedge in affected games and replaying the observed rating rule. The dashed vertical line marks the sample mean.

Figure 9: K schedules shift incorporation bias and learning speed



Notes: Panel A reports incorporation bias, measured as the mean counterfactual rating gain from adding back the estimated attenuation wedge in affected games. Panel B reports learning speed, measured as the median number of games required for a player's rating percentile to reach sustained convergence within five percentage points of the player's latent-ability percentile. Schedules are sorted by incorporation bias. Lower values are preferred on both dimensions.

Figure 10: Equity–efficiency frontier across convergence thresholds



Notes: Each panel plots counterfactual update schedules under a different sustained-convergence benchmark. The horizontal axis reports median games to convergence; the vertical axis reports incorporation bias, measured as the mean attenuation rating penalty. Orange points identify non-dominated schedules on the lower-left frontier. The red star marks the current K rule. Lower-left is preferred: faster learning and smaller incorporated attenuation penalty.

Table 1: Key USCF rating-rule parameters

Object	Institutional meaning	Role in the paper
S_{ie}	Total event score, obtained by summing wins, draws, and losses.	Objective realized performance.
E_{ie}	Total expected event score implied by pregame ratings and opponent ratings.	Benchmark for rating-system expectations.
K_{ie}	Rating sensitivity parameter; larger values convert score deviations into larger rating changes.	Governs how quickly performance signals become ratings.
B_t	Calendar-time bonus threshold parameter.	Determines how hard it is to trigger extra rating points.
m'_{ie}	Event-length adjustment, equal to $\max\{m_{ie}, 4\}$.	Scales the bonus threshold by event size.
Bonus term	Extra rating points awarded when $K_{ie}(S_{ie} - E_{ie})$ exceeds $B_t\sqrt{m'_{ie}}$.	Creates bonus-relevant high-stakes states.

Table 2: Female-focal analytical sample summary

	Overall	Female opponent	Male opponent
Panel A. Sample size and coverage			
Player-games	978,135	173,486	804,649
Share of games	100.0%	17.7%	82.3%
Female focal players	8,037	7,374	7,981
Unique opponents	–	23,847	155,896
Events	71,585	28,111	70,725
Event-sections	114,226	39,497	112,346
Sample period	1994–2023	1994–2023	1994–2023
Panel B. Game outcomes			
Mean score	0.526	0.597	0.510
Win share	46.7%	54.4%	45.1%
Draw share	11.6%	10.7%	11.8%
Loss share	41.6%	34.9%	43.1%
Panel C. Ratings and demographics			
Mean focal rating	1,287.1	1,245.5	1,296.1
Mean opponent rating	1,245.1	1,119.4	1,272.2
Mean rating gap	42.0	126.1	23.9
Mean absolute rating gap	252.1	288.3	244.3
Mean focal age	12.86	12.44	12.95
Mean opponent age	19.44	13.84	20.65

Notes: The table summarizes the main female-focal analytical sample. Each observation is a female focal player-game. Game scores equal one for wins, one-half for draws, and zero for losses. The opponent-gender columns split the sample by whether the opponent is female or male. Player, opponent, event, and event-section counts are not additive across opponent-gender columns because the same player, opponent, event, or section can appear in both columns. Ratings are pre-game ratings.

Table 3: Near-threshold score response and male-opponent attenuation

Dependent variable: result score	(1)	(2)	(3)	(4)	(5)	(6)
Near-below cutoff, 1K	0.0118***	0.0152***	0.0056**	0.0009	0.0065*	0.0116***
	(0.0020)	(0.0020)	(0.0027)	(0.0028)	(0.0035)	(0.0036)
Male opponent	-0.0126***	-0.0156***	-0.0157***	-0.0217***	-0.0169***	-0.0203***
	(0.0015)	(0.0016)	(0.0024)	(0.0024)	(0.0033)	(0.0033)
Near-below cutoff × Male opponent	-0.0100***	-0.0096***	-0.0102***	-0.0080***	-0.0096**	-0.0098**
	(0.0021)	(0.0021)	(0.0030)	(0.0030)	(0.0039)	(0.0039)
Player FE	No	Yes	No	Yes	No	Yes
Event-section-round FE	No	No	Yes	Yes	No	No
Score-cell FE	No	No	No	No	Yes	Yes
Observations	978,135	978,135	978,135	978,135	978,135	978,135
R^2	0.246	0.254	0.582	0.590	0.734	0.744

Notes: This table reports reduced-form estimates for female focal players. The dependent variable is the game result score. "Near-below cutoff, 1K" indicates that the focal player is within one K-unit below the bonus-relevant rating threshold. "Male opponent" indicates that the opponent is male. The interaction term measures how the near-threshold response differs in male-opponent games relative to female-opponent games. Column (6) is the preferred specification. Standard errors are reported in parentheses. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4: Robustness: alternative comparison groups

Dependent variable: result score	Main	Below-only	Donut	Local $[-2K, .5K]$	Local below
Near-below cutoff $[-1K, 0)$	0.0116***	0.0107***	0.0089**	0.0138***	0.0119***
	(0.0036)	(0.0039)	(0.0038)	(0.0038)	(0.0041)
Male opponent	-0.0203***	-0.0216***	-0.0217***	-0.0203***	-0.0218***
	(0.0033)	(0.0037)	(0.0036)	(0.0036)	(0.0040)
Near-below cutoff × Male opponent	-0.0098**	-0.0087**	-0.0088**	-0.0104**	-0.0095**
	(0.0039)	(0.0042)	(0.0041)	(0.0041)	(0.0044)
Player FE	Yes	Yes	Yes	Yes	Yes
Score-cell FE	Yes	Yes	Yes	Yes	Yes
Observations	978,135	865,550	897,907	910,899	830,671
R^2	0.744	0.748	0.749	0.749	0.753
Within R^2	0.167	0.175	0.172	0.171	0.176

Notes: This table reports robustness checks using alternative comparison groups. The sample consists of female focal-player games. "Main" uses the baseline comparison window. "Below-only" restricts the comparison group to games below the threshold. "Donut" excludes observations closest to the cutoff outside the treated window. "Local $[-2K, .5K]$" restricts the comparison to a more local window around the cutoff, and "Local below" restricts the comparison to local below-threshold games. All specifications include player fixed effects and score-cell fixed effects. Standard errors are reported in parentheses. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: Implied attenuation across alternative comparison groups

Specification	$\widehat{\beta}_1$	$\widehat{\beta}_1 + \widehat{\beta}_3$	$\widehat{\rho}$	Attenuation $1 - \widehat{\rho}$
Main	0.0116	0.0018	0.155	0.845
Below-only	0.0107	0.0020	0.187	0.813
Donut	0.0089	0.0001	0.011	0.989
Local $[-2K, \ldots]$	0.0138	0.0034	0.246	0.754
Local below	0.0119	0.0024	0.202	0.798

Notes: This table converts the robustness estimates in Table 4 into attenuation metrics. $\widehat{\beta}_1$ is the estimated near-threshold response against female opponents. $\widehat{\beta}_1 + \widehat{\beta}_3$ is the corresponding response against male opponents. $\widehat{\rho}$ is the share of the near-threshold response that remains in male-opponent games. Attenuation $1 - \widehat{\rho}$ is the implied missing share of the response. Values are computed from rounded coefficients.

Table 6: Player-level incorporation gaps by exposure

		Incorporation Gap (Elo)		
Exposure group	Players	Mean	Median	P90
All players	**8,037**	**15.56**	**7.88**	**41.48**
0 affected games	152	0.00	0.00	0.00
1–5 affected games	1,212	1.02	1.01	1.88
6–20 affected games	2,039	3.57	3.34	5.91
>20 affected games	4,634	25.15	17.59	54.71

Notes: The table reports player-level cumulative counterfactual rating gains, in Elo points, from replaying the observed rating rule after adding back the estimated score-unit attenuation wedge in affected games. The top row reports the aggregate distribution (where 98.1% of players were affected in at least one game), while subsequent rows decompose the gaps by the number of affected-game exposures.

Table 7: Stock-flow decomposition of replay-implied score losses

Sample	Rating-absorbed share	Direct current-game share
All male-opponent games	87.5%	12.5%
Affected games only	79.7%	20.3%

Notes: The rating-absorbed share is the fraction of the replay-implied score loss that is already reflected in the player's rating before the current game begins. The direct current-game share is the fraction attributable to the contemporaneous score wedge in the current game.

Table 8: Counterfactual K-update schedules

Schedule family	Examples	Design interpretation
Status quo	Observed USCF K	Benchmark rule: high early updating, lower later updating.
Scaled status quo	Current rule $\times 75\%$, $\times 50\%$, $\times 25\%$	Preserves the shape of the existing rule while reducing overall update speed.
Flat rules	$K = 8, 12, 16, 24, 32$	Makes the update rate independent of player experience and rating.
High-then-low rules	$32 \to 16$ after 25 games; $32 \to 24$ after 25 games	Learns aggressively from early games, then reduces pass-through later.
Low-then-high rules	$8 \to 24$ after 25 games; $12 \to 24$ after 25 games; $12 \to 32$ after 50 games	Insures early ratings against noisy or biased signals, then allows faster later adjustment.
Ramp rules	Ramp up $8 \to 24$ over 50 games; ramp down $32 \to 16$ over 50 games	Smooths the transition between low and high update regimes.
Caps and floors	Observed K capped at 16 or 24; observed K floored at 16	Limits extreme update rates while preserving part of the current rule.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Boston: Harvard Business Review Press.

Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). (2019). *The economics of artificial intelligence: An agenda*. Chicago: University of Chicago Press.

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, *30*(2), 175–187.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–42). Princeton: Princeton University Press.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Backus, P., Cubel, M., Guid, M., Sánchez-Pagés, S., & López Mañas, E. (2023). Gender, competition, and performance: Evidence from chess players. *Quantitative Economics*, *14*(1), 349–380. doi: 10.3982/QE1404

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, *100*(3), 598–614.

Becker, G. S. (1957). *The economics of discrimination*. Chicago: University of Chicago Press.

Belgian Go Federation. (2024). *Belgian go federation rating system.* (Accessed 2026-07-06)

Benson, A., Li, D., & Shue, K. (2019). Promotions and the peter principle. *The Quarterly Journal of Economics*, *134*(4), 2085–2134. doi: 10.1093/qje/qjz022

Bertrand, M., & Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. doi: 10.1257/0002828042002561

Bilalić, M., Smallbone, K., McLeod, P., & Gobet, F. (2009). Why are the best women so good at chess? participation rates and gender differences in intellectual domains. *Proceedings of the Royal Society B*, *276*(1659), 1161–1165.

Bohren, J. A., Hull, P., & Imas, A. (2025). Systemic discrimination: Theory and measurement. *Quarterly Journal of Economics*, *140*(3), 1743–1799.

Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, *109*(10), 3395–3436.

Brancaccio, A., & Gobet, F. (2023). Scientific explanations of the performance gender gap in chess and science, technology, engineering and mathematics. *Journal of Expertise*, *6*(1).

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, *119*(5), 982–1013.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, *129*(3), 1409–1447. doi: 10.1093/qje/qju009

Caselli, M., Falco, P., & Mattera, G. (2023). When the stadium goes silent: How crowds affect the performance of discriminated groups. *Journal of Labor Economics*, *41*(2), 431–451. doi: 10.1086/719967

Chabris, C. F., & Glickman, M. E. (2006). Sex differences in intellectual performance: Analysis of a large cohort of competitive chess players. *Psychological Science*, *17*(12), 1040–1046.

Chevalier, J., & Ellison, G. (1997). Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, *105*(6), 1167–1200. doi: 10.1086/516389

Coate, S., & Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, *83*(5), 1220–1240.

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? or biased data? a field experiment in operationalizing ai ethics. *Proceedings of the 21st ACM Conference on Economics and Computation*.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.

De Sousa, J., & Hollard, G. (2023). From micro to macro gender differences: Evidence from field tournaments. *Management Science*, *69*(6), 3358–3399. doi: 10.1287/mnsc.2022.4541

Düring, B., Fischer, M., & Wolfram, M.-T. (2022). An elo-type rating model for players and teams of variable strength. *Philosophical Transactions of the Royal Society A*, *380*(2224), 20210155.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. New York: Arco Publishing.

Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*, *138*(4), 2021–2067. doi: 10.1093/qje/qjad025

Exley, C. L., & Kessler, J. B. (2022). The gender gap in self-promotion. *Quarterly Journal of Economics*, *137*(3), 1345–1381.

Gerdes, C., & Gränsmark, P. (2010). Strategic behavior across gender: A comparison of female and male expert chess players. *Labour Economics*, *17*(5), 766–775.

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, *28*(6), 673–689.

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *Quarterly Journal of Economics*, *132*(3), 1219–1260.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, *118*(3), 1049–1074.

Healy, P. M. (1985). The effect of bonus schemes on accounting decisions. *Journal of Accounting and Economics*, *7*(1–3), 85–107. doi: 10.1016/0165-4101(85)90029-1

Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, *10*(1), 74–91.

Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, *7*, 24–52.

Hvattum, L. M., & Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470.

Kessler, J. B., Low, C., & Sullivan, C. D. (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, *109*(11), 3713–3744. doi: 10.1257/aer.20181714

Kisida, B., Pepper, M., Podgursky, M., & Wickman, M. (2026). Searching for the queen's gambit: An analysis of male-female performance gaps in us chess. *Journal of Sports Economics*, 15270025261438415.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, *108*, 22–27. doi: 10.1257/pandp.20181018

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Kovalchik, S. (2020). Extension of the elo rating system to margin of victory. *International Journal of Forecasting*, *36*(4), 1329–1341.

Kovalchik, S. A. (2016). Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138.

Larkin, I. (2014). The cost of high-powered incentives: Employee gaming in enterprise software sales. *Journal of Labor Economics*, *32*(2), 199–227. doi: 10.1086/673371

Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, *89*(5), 841–864.

Maass, A., D'Ettole, C., & Cadinu, M. (2008). Checkmate? the role of gender stereotypes in the ultimate intellectual sport. *European Journal of Social Psychology*, *38*(2), 231–245. doi: 10.1002/ejsp.440

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. doi: 10.1257/jep.31.2.87

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics*, *122*(3), 1067–1101.

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*, 601–630. doi: 10.1146/annurev-economics-111809-125122

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. doi: 10.1126/science.aax2342

Oyer, P. (1998). Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *The Quarterly Journal of Economics*, *113*(1), 149–185. doi: 10.1162/003355398555559

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, *101*(4), 1410–1435. doi: 10.1257/aer.101.4.1410

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*(4), 659–661.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7–63.

Rothgerber, H., & Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group Processes & Intergroup Relations*, *17*(1), 79–90.

Smerdon, D., Hu, H., McLennan, A., von Hippel, W., & Albrecht, S. (2020). Female chess players show typical stereotype-threat effects: Commentary on stafford (2018). *Psychological Science*, *31*(6), 756–759. doi: 10.1177/0956797620924051

Stafford, T. (2018). Female chess players outperform expectations when playing men. *Psychological Science*, *29*(3), 429–436. doi: 10.1177/0956797617736887