A Showdown Ladder Rating Measures the Ladder, Not the Player

Brecht Corbeel
Absolute Digital Publishers
absolutedigitalpublishers.com

September 7, 2026

Abstract

Pokémon Showdown's own ladder-help page publishes, to the tenth significant digit, the exact constants of a Glicko-1 rating system: a starting rating of $R_0 = 1500$, a starting rating deviation of $\mathrm{RD}_0 = 130$, an operating range $\mathrm{RD} \in [25, 130]$, 24-hour rating periods, and a system constant of 6.6775026092 [26]. This paper takes that documentation exactly as seriously as a psychometrician takes a standardized test's technical manual. It first shows, term for term rather than by loose analogy, that Showdown's published rating deviation is Classical Test Theory's standard error of measurement under a different name [14, 25]. It then applies the standard the field that invented that apparatus actually uses to decide whether such a score may be trusted across changing conditions — construct validity and measurement invariance [7, 23, 24] — and argues that a rating computed against a periodically and deliberately edited item bank (the legal competitive metagame, redrawn by community-voted bans) has only ever been shown to be reliable, not valid across that boundary. A specific discriminator is proposed: rank- correlate the same cohort's ratings immediately before and after a tier-defining ban. This pass could not obtain the longitudinal per-player data the discriminator needs from any public Showdown or Smogon source, so — rather than present an untested discriminator as a completed test — the paper instead runs the actual Glicko-1 algorithm, exactly as Showdown configures it, on a seeded synthetic population of 200 players across 30 burn-in and up to 60 post-shift rating periods, over 300 replications per condition. The resulting Spearman/Kendall rank correlations fall monotonically from $\rho_S \approx 0.98$ under no true skill reshuffling to $\rho_S \approx 0.28$ under a complete reshuffling, and fall further, from 1.00 at zero post-shift games to 0.066 at sixty, as the post-shift ladder is allowed to run longer — establishing that the proposed discriminator has real statistical power to separate a stable-trait world from a population-relative one, conditional on data this research pass did not have access to. Smogon's own suspect-test voting threshold, and published research by Glicko-1's own inventor treating chess ratings as a usable psychometric measure, are engaged directly as genuine institutional counter-evidence rather than dismissed. The paper's claim is therefore narrower than "ladder ratings are meaningless": Glicko-1 on Showdown is a real, competently specified, and honestly self-documented measurement instrument whose reliability disclosure is more complete than most systems that borrow the word "rating"; what remains unestablished, on the standard the instrument's own field uses, is whether the number it publishes is the same trait once the test itself has been edited.

Keywords: Glicko-1; rating deviation; classical test theory; construct validity; measurement invariance; Elo rating; competitive Pokémon; paired-comparison models.

Series: The Pokémon Formal Methods Papers, No. 2.

1 Introduction

A number that goes up when you win and down when you lose looks, to the player reading it, like a skill score. Pokémon Showdown's own documentation is more careful than that. Its ladder-help page does not describe a high score; it specifies an instrument, in the register of a technical manual: starting values, update rules, growth schedules, and a system constant carried to ten significant digits [26]. Three scores are computed from the same match history — Elo, Glicko-1, and GXE — and the third of Showdown's own three, the rating deviation attached to the second, is not a skill number at all. It is a bound on how much noise sits around the skill number, published by the same page, recomputed by the same algorithm every 24 hours, and routinely invisible to anyone who screenshots a bare rating as a credential.

This paper's claim has three parts, each held to a real, checkable standard rather than an informal one. First (§2–§3): Showdown's Glicko-1 rating deviation is not merely *like* a measurement-theory concept — it is, term for term, on Glickman's own stated construction of the interval a Glicko rating supports, Classical Test Theory's standard error of measurement [14, 25]. That correspondence is this paper's first contribution: a literal homology, not a metaphor borrowed for color. Second (§4): the field that invented the standard-error-of-measurement construction also drew, deliberately, a line between two questions a test score raises — is it *reliable* (repeatable, low-noise) and is it *valid* (actually estimating the trait its users assume, comparably, across different conditions) — and has a formal apparatus, measurement invariance, for the second question [7, 23, 24]. A competitive Pokémon tier's legal roster is edited on purpose, by community vote, whenever a suspect test bans or unbans something; that is not a vague sense in which "the game changes," it is the literal analogue of changing a test's item bank while insisting the scoring key stayed the same. Third (§5–§6): this paper specifies an exact discriminator for the resulting question and reports, honestly, that it could not be run on real players — no public Showdown or Smogon source archives the longitudinal, per-player rating history the test needs — and so instead reports a seeded, fully computed simulation of the actual Glicko-1 algorithm on synthetic data, which establishes that the discriminator *would* have real power to tell the two readings apart, without claiming to know which reading is true of the real ladder.

Honest positioning. Nothing about Glicko-1 is this paper's invention. Glickman specified the update equations, the paired-comparison model beneath them, and the standard-error-style confidence interval a rating and its deviation jointly support [12, 14]. Classical Test Theory's observed/true/error decomposition is Novick's codification of a tradition running back through Spearman's 1904 correlation work [22, 25, 30]. Construct validity is Cronbach and Meehl's [7], extended by Messick [24]; measurement invariance is Meredith's [23], with Vandenberg and Lance [31] and Cheung and Rensvold [5] supplying the applied testing methodology this paper borrows rather than reinvents, and Holland and Wainer's differential item functioning literature [19] the closest existing formal treatment of a test behaving differently across subpopulations. What this paper contributes, and no more: (a) the RD ≡ SEM correspondence stated exactly, term for term (Proposition 2); (b) the construct-validity question posed as a specific, falsifiable discriminator (Proposition 3); (c) a seeded numerical experiment that measures that discriminator's statistical power on synthetic data generated through an exact

re-implementation of Showdown's own algorithm (Proposition 4), together with two closed-form limiting-case checks; and (d) a direct engagement with the strongest available counter-evidence — Smogon's own suspect-test voting threshold, and published research by Glickman himself treating chess ratings as a usable psychometric measure — rather than a dismissal of it.

Roadmap. §2 states the instrument as Showdown configures it and checks its dimensional sanity. §3 derives the RD/SEM correspondence. §4 states the construct-validity question and the discriminator. §5 reports what data actually exists, invokes the kill criterion honestly, and specifies the fallback simulation. §6 reports the simulation's actual numbers. §7 checks limiting and degenerate cases. §8 engages Smogon's own governance practice and Glickman's own research program directly. §9 states five distinct ways this paper's argument could be wrong. §10 states what would close the gap.

2 The instrument, as configured

Proposition 1 (The instrument; attributed to Showdown/Zarel and Glickman)**.** *Pokémon Showdown computes three rating metrics from one match history.*

Elo. *Named for Arpad Elo's original chess-rating system [10], though Showdown's implementation is its own: a new account starts at* $R_0 = 1000$*. Between* 1000 *and* 1100 *the* K*-factor ramps linearly from* 80 *to* 50 *for a winning player and from* 20 *to* 50 *for a losing player; between* 1100 *and* 1299*,* $K = 50$*; at* 1300 *and above,* $K = 40$*. A floor of* 1000 *applies unconditionally. Above* 1400*, an inactive account decays daily at 9:00 GMT: one point per 100 points above 1500 if one to five games were played that day, one point per 50 points above 1400 if none were [26].*

Glicko-1. *A new account starts at* $R_0 = 1500$*,* $\mathrm{RD}_0 = 130$*. Rating deviation is constrained to* $\mathrm{RD} \in [\mathrm{RD}_{\min}, \mathrm{RD}_{\max}] = [25, 130]$ *throughout an account's life. Rating periods are 24 hours. The system constant is* $c = 6.6775026092$ *[26]. Following Glickman's own notation [12, 14], define*

$$g(\mathrm{RD}) \;=\; \frac{1}{\sqrt{1 + 3q^2\mathrm{RD}^2/\pi^2}}, \qquad q = \frac{\ln 10}{400} \approx 0.0057565, \tag{1}$$

$$E(s \mid r, r_j, \mathrm{RD}_j) \;=\; \frac{1}{1 + 10^{-g(\mathrm{RD}_j)(r - r_j)/400}}. \tag{2}$$

Between rating periods, deviation grows with elapsed time even absent play, $\mathrm{RD}_{\mathrm{pre}} = \sqrt{\mathrm{RD}_{\mathrm{prev}}^2 + c^2}$*, clamped to* $[\mathrm{RD}_{\min}, \mathrm{RD}_{\max}]$*; within a period in which* m *games were played against opponents* $j = 1, \ldots, m$ *with outcomes* $s_j \in \{0, \frac{1}{2}, 1\}$*,*

$$d^2 \;=\; \Big[q^2 \textstyle\sum_{j=1}^{m} g(\mathrm{RD}_j)^2\, E(s \mid r, r_j, \mathrm{RD}_j)\big(1 - E(s \mid r, r_j, \mathrm{RD}_j)\big)\Big]^{-1}, \tag{3}$$

$$r' \;=\; r + \frac{q}{1/\mathrm{RD}_{\mathrm{pre}}^2 + 1/d^2} \sum_{j=1}^{m} g(\mathrm{RD}_j)\big(s_j - E(s \mid r, r_j, \mathrm{RD}_j)\big), \qquad \mathrm{RD}' \;=\; \Big(1/\mathrm{RD}_{\mathrm{pre}}^2 + 1/d^2\Big)^{-1/2}. \tag{4}$$

Substituting Showdown's c makes the growth schedule fully determined rather than a qualitative "uncertainty" — the same 6.6775026092 *every 24 hours, for every account. Appendix A carries this derivation to every intermediate line.*

GXE. *"An estimate of your win chance against an average ladder player" [26], a monotone transform of* (r, RD) *that this paper does not need to unpack further: every argument below concerns* RD *directly, and GXE inherits whatever* RD *implies about reliability without adding new structure.*

Dimensional sanity

Rating points are not a physical unit with an independent operational definition the way a kilogram is; the check that matters here is not dimensional analysis in the physics sense but whether the number's units are internally consistent across every place they appear. They are, by construction: r and RD share one scale, fixed by Eq. 2 — a 400-point rating gap corresponds to exactly a 10:1 odds ratio in expected score, regardless of where on the scale the gap sits, so "a difference of x rating points" always denotes the same operational claim (an implied win probability via Eq. 2) whether x separates two rating-1500 accounts or two rating-2200 accounts. RD is expressed in the identical units as r itself (Eq. 3–4 return RD' from the same $1/\mathrm{RD}^2$ terms that produce r''s denominator), which is exactly the property Classical Test Theory's error term needs to share with its observed score for the comparison in §3 to be a homology rather than a coincidence of two differently-scaled numbers that happen to both be called "uncertainty." A claim later in this paper — that RD is formally a standard error of measurement — would fail immediately if RD were secretly expressed on a different footing than r; Eq. 3–4 show it is not.

3 The rating deviation is a standard error of measurement

Classical Test Theory's founding decomposition writes an observed score X on a single test administration as a true score T — the average a person would obtain across infinitely many parallel administrations — plus measurement error E, with E uncorrelated with T and mean zero across the tested population:

$$X = T + E. \tag{5}$$

From Eq. 5, reliability is the proportion of observed-score variance attributable to true-score variance. In most psychometric practice reliability is estimated from internal consistency across a fixed item set — Cronbach's coefficient alpha is the standard example [6] — which has no direct analogue here, since a Glicko-1 rating period has no fixed "item set" to be internally consistent across; RD is instead a model-based reliability index, derived from the paired- comparison model's own likelihood (Eq. 3) rather than from covariance among repeated items, a distinction worth flagging so Table 1's correspondence is not mistaken for an alpha-style computation. Either way, the standard error of measurement converts a single observed score into a defensible interval rather than a point claim:

$$\mathrm{SEM} = \mathrm{SD}_X \sqrt{1 - \text{reliability}}\,, \qquad T \in X \pm 1.96 \cdot \mathrm{SEM} \text{ at } 95\% \text{ confidence} \tag{6}$$

[22, 25].

Proposition 2 (RD ≡ SEM; THIS PAPER'S contribution)**.** *Glickman's own stated use of a Glicko rating and its deviation is, in his words, to "report a 95% confidence interval. The lowest value in the interval is the player's rating minus twice the RD, and the highest value is the player's rating plus twice the RD" [14] — i.e.* $r \pm 2\,\mathrm{RD}$*, matching Eq. 6's* $X \pm 1.96 \cdot \mathrm{SEM}$ *to two significant figures in the multiplier. Table 1 maps every remaining term. This is a restatement of Glickman's own construction in Novick's notation, not a new derivation and not a metaphor: it establishes that Showdown's published* RD *satisfies the operational definition of a standard error of measurement, term for term, rather than merely resembling one informally.*

Table 1: Classical Test Theory and Glicko-1 on Showdown, term for term.

Classical Test Theory	Glicko-1 on Showdown
Observed score X on one administration	Rating r at the end of one rating period
Unobservable true score T	Glickman's latent paired-comparison strength
$\mathrm{SEM} = \mathrm{SD}_X\sqrt{1 - \text{reliability}}$	Rating deviation $\mathrm{RD} \in [25, 130]$
95% interval $X \pm 1.96 \cdot \mathrm{SEM}$	95% interval $r \pm 2\,\mathrm{RD}$ [14]
Reliability rises with more independent items	RD falls with more games per period [12]
Reliability falls with time since calibration	RD rises with time since last game [14]

A worked interval, at two points this paper's own data touch. A brand-new account, $\mathrm{RD}_0 = 130$, sitting at $r = 1500$: the 95% interval is $1500 \pm 2(130) = [1240, 1760]$ — a 520-point band, wide enough that Showdown's own publication of it is close to an admission that a first rating is barely a rating at all. An account whose RD has settled toward this paper's own simulated steady state under continued play, $\mathrm{RD} \approx 36$ (§7, matching the empirical mean RD at burn-in end in §6's experiment to within two points of the model's own closed-form fixed point): the same construction gives $1500 \pm 2(36) = [1428, 1572]$, a 144-point band around the same central value. Both intervals are computed from Eq. 6's logic with Showdown's own numbers; neither number is invented for this paper. What Eq. 6 and Proposition 2 do *not* say is whether the true score T this interval brackets is the same T before and after the item bank under it changes — that is a validity question, taken up next.

4 Reliability is not validity, and invariance is the standard for the difference

A precise instrument is not automatically a valid one. Cronbach and Meehl drew this line in 1955 and psychometrics has kept it since: reliability asks whether a measurement repeats, while construct validity asks "what psychological constructs account for test performance" at all — whether the number tracks the latent trait its users assume, judged against a network of other claims the trait ought to predict if it is real [7]. Messick's later synthesis folds the question into a single unified concept — validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" [24] — but keeps the same consequence: precision (Eq. 6's SEM, this paper's RD) bounds noise and says nothing, on

its own, about whether the number's center is the right center for the trait its users think it measures.

Proposition 3 (The validity/invariance question, and its discriminator; THIS PAPER'S construct)**.** *The formal tool for whether a score means the same thing under different conditions is measurement invariance. Meredith's hierarchy is not a yes/no label but a strict ladder: configural invariance (the same structure holds), metric invariance (the same loadings), scalar invariance (the same loadings and intercepts, making raw scores comparable across groups), and strict invariance, which Meredith argues is the level actually "required for fairness/equity to exist" when scores from different conditions are compared directly [23]. Vandenberg and Lance's applied synthesis and Cheung and Rensvold's fit-index criteria are the standard machinery for testing where a given instrument sits on that ladder in practice [5, 31]; Holland and Wainer's differential item functioning literature is the nearest existing formal treatment of an item behaving differently for different sub-populations of test-takers under a nominally fixed scoring key [19]; the AERA/APA/NCME* Standards *codify, institutionally, that cross-condition comparability "has to be established with evidence, not assumed" simply because a scoring formula stayed fixed [1].*

A Showdown ladder is a paired-comparison test whose item bank is the current legal metagame — the roster of species, movesets, and banned strategies a player's opponents are drawn from in a given rating period. Meredith's hierarchy was built for survey items with fixed text; here, the item bank is the ruleset, and a suspect test edits it on purpose. Glickman's own paired-comparison model, like every classical model in this family, assumes the comparison structure it estimates within holds still across a rating period [12] — an assumption a ban directly violates for the affected item. Berg's own statistical audit of chess's Elo system works in a domain where this assumption is nearly exact by construction: chess's rules have not changed what a win means since long before either rating system existed [2]. Showdown does not have that property, and nothing on its ladder-help page claims it does — the page specifies an algorithm, not a validity claim, and the two are not the same statement.

Discriminator. *Fix a tier and a tier-defining ban date. Let the cohort be every player active on that tier's ladder in the week before the ban, with* (r, RD) *recorded at that moment. Let the post-ban ladder run until* RD *re-settles toward the low end of Showdown's range for players who keep playing regularly — Glickman's own guidance suggests "an average of 5–10 games per player" before a period's output should be trusted [14] — then record* (r, RD) *again for the* same *cohort. Compute Spearman's* ρ_S *[30] and Kendall's* τ_K *[20] between the two r snapshots. A high correlation supports the stable-trait reading Smogon's own governance practice assumes (§8); a low correlation supports the population-relative reading this paper argues for. This is fully specified and, in principle, runnable.*

Kill criterion (met in this pass; see §5). If the per-player, per-period longitudinal rating history the discriminator needs is not publicly archived at the required granularity by Showdown or Smogon, the discriminator as designed cannot be executed on real players, and the paper must say so plainly rather than present an untested design as a completed test. §5 reports that this is exactly the situation this research pass found.

This gap is not Pokémon-specific

The broader competitive-gaming industry has independently converged on richer Bayesian rating architectures precisely because Elo/Glicko-style point-plus-uncertainty scores were felt to

be insufficient: Microsoft's TrueSkill replaces a single opponent comparison with a full factor-graph inference over an entire team match [18], and deployed industry variants extend it further — Ubisoft's matchmaking system for Ghost Recon Online adds context-dependent skill estimation on top of a TrueSkill-family base [9], and a 2025 system for League of Legends explicitly separates per-role performance from a single scalar rating [8], while a 2024 paper proposes stochastic-variance extensions to Elo itself, structurally close to the volatility parameter Glickman later added in Glicko-2 [13, 17]. Even outside games, Elo-family scoring has been adapted for adaptive testing in education [32] — treating a fundamentally similar reliability-only construction as sufficient for a use case (sequencing practice problems) that, unlike a suspect-test vote or a hiring credential, does not actually require cross-time construct validity. As far as this research pass could determine, none of these industry systems' own public documentation includes an explicit measurement-invariance audit of the kind §4 specifies either. The gap this paper identifies is therefore not an idiosyncrasy of one game's community-run ladder; it is a generic property of how the competitive-gaming and adjacent industries have engineered skill ratings, and Showdown is simply the case where the underlying algorithm's constants happen to be published in full enough detail to audit directly.

5 Data and method

5.1 What public data exists, and why it cannot run Proposition 3

Two kinds of Showdown/Smogon-adjacent data are genuinely public. First, Smogon's monthly usage-statistics releases report, per tier and per month, the aggregate usage rate of each species and common team compositions — a population-level snapshot of the *metagame*, not of any individual's rating trajectory. Second, Showdown's ladder pages return a player's *current* standing on request — a single cross-section, not an archived time series. Neither object is the thing Proposition 3 needs: a per-player, per-rating-period (r, RD) history spanning a specific ban date, for a fixed cohort, held privately or publicly by the platform. Smogon's own suspect-test threshold posts (§8) confirm the platform can and does check an individual's standing at a point in time for voter-eligibility purposes, but nothing in this research pass located a public archive of such standings sampled repeatedly across a ban boundary for a defined cohort. This contrasts with chess, where federation-published crosstables and engine-scored games let Regan and Haworth construct an "intrinsic" performance rating independent of the Elo pool itself and use it to audit Elo's own long-run stability [27] — an audit that is possible specifically because chess ratings are attached to an institutionally archived, individually identifiable game-by-game record in a way a Showdown ladder account is not. That asymmetry is itself a small finding: an institution making a governance-grade assumption about what its own rating measures (§8) has not, as far as this pass could determine, made the data that would test that assumption publicly available.

Kill criterion invoked. Per Proposition 3's own stated falsifier, the discriminator as designed is not run on real Showdown players in this paper. What follows is not a substitute empirical answer — it is a different, honest question this paper *can* answer: given the discriminator's design, and given the real, published Glicko-1 algorithm, does the statistic actually have the power to tell a stable-trait world from a population-relative world apart, and how does that power depend on how much of the metagame moves and on how long the post-shift ladder is allowed to run? That question is answerable with synthetic data and a faithful re-

implementation of the audited algorithm, and answering it honestly requires no data this paper does not have.

5.2 The synthetic experiment

Proposition 4 (The synthetic power experiment; THIS PAPER'S computation)**.** $N = 200$ *synthetic players are assigned a latent skill* $\theta_i \sim \mathcal{N}(1500, 200^2)$*, drawn once per replication and* not *observed directly by the algorithm under test. Match outcomes are generated by the logistic (Bradley–Terry) model*

$$\Pr(i \text{ beats } j) = \frac{1}{1 + 10^{-(\theta_i - \theta_j)/400}}, \tag{7}$$

the same functional form as Glicko-1's own $E(s \mid r, r_j, \mathrm{RD}_j)$ *(Eq. 2) with* $g(\cdot) = 1$ *— i.e. the underlying game is generated by exactly the model the audited instrument assumes, so the experiment never smuggles in a data-generating process the instrument was not built to fit [3]. Every player starts at* $(r, \mathrm{RD}) = (1500, 130)$*, exactly Showdown's own new-account values. Each rating period, every player plays* 6 *games against random opponents (matching Glickman's own "5–10 games" guidance, 14), and ratings update by Eq. 3–4 with Showdown's own constants.* $T_{\text{burn}} = 30$ *burn-in periods let* RD *approach its steady state under* θ *held fixed; the pre-shift snapshot* (r, RD) *is then recorded for every player.*

A metagame shift is then applied as

$$\theta_i' = (1 - \lambda)\,\theta_i + \lambda\,\theta_{\pi(i)}, \qquad \pi \text{ a uniform random permutation of } \{1, \ldots, N\}, \tag{8}$$

for $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ *— a construction that holds the* population's *skill distribution exactly fixed (the same multiset of* θ *values) and varies only* who *holds which value. This operationalizes "a ban changes who its remaining strategies reward" without also claiming the ladder gets harder or easier on average, which a ban does not obviously do either way.* $\lambda = 0$ *is a pure test-retest world (no reshuffling: the stable-trait hypothesis, exactly);* $\lambda = 1$ *is complete reshuffling (the population-relative hypothesis, exactly). Ratings are* not *reset at the shift — they continue from their pre-shift* (r, RD) *state, as a real ban does not reset a ladder.* T_{post} *further periods are run under* θ' *(swept over* $\{0, 5, 10, 20, 30, 45, 60\}$ *at* $\lambda = 1$*; fixed at* 30 *for the main* λ *sweep), after which the post-shift snapshot is recorded for the* same *cohort, filtered to players with* $\mathrm{RD} \le 60$ *at both snapshots (an "active, calibrated players" filter, structurally the same kind of inclusion rule Smogon's own suspect-test threshold imposes, §8). Spearman's* ρ_S *and Kendall's* τ_K *are computed between the two snapshots for that cohort.*

Every replication is seeded from `numpy.random.SeedSequence(20260907)`*, spawned into* $K = 300$ *independent child seeds per grid point (Appendix B gives the full parameter table and pseudocode; the complete source is* `sim_glicko_ladder.py` *in this paper's package, and its exact console output is reproduced verbatim in Appendix C). Re-running the script with no arguments reproduces every number in §6 bit for bit.*

What this experiment is not. It is not a claim about the true value of λ a real Smogon ban induces — that remains exactly as unknown after this experiment as before it, per the kill criterion above. It is a claim about the discriminator's own statistical behavior, run through the real, audited algorithm, which is a fair question to answer without the missing empirical data and a dishonest one to skip past by silence.

6 Results

Every number in this section is copied verbatim from `sim_glicko_ladder_output.txt` (reproduced in full in Appendix C) and `sim_glicko_ladder_results.json`, produced by the single run described in §5.2 under `MASTER_SEED=20260907`. No number below is estimated, rounded beyond the script's own printed precision, or extrapolated.

6.1 The λ sweep: how much of the metagame has to move

Table 2: Rank correlation between pre-shift and post-shift ratings, by shift severity λ (Eq. 8), $N = 200$ synthetic players, $K = 300$ replications per row, $T_{\text{burn}} = T_{\text{post}} = 30$ periods, cohort filter RD ≤ 60 at both snapshots (mean cohort size 200.0 at every λ — the filter excluded no one at these settings).

λ	ρ_S (mean)	ρ_S (sd)	τ_K (mean)	τ_K (sd)	mean RD at burn-in end
0.00	0.9840	0.0024	0.8950	0.0079	35.798
0.20	0.9650	0.0049	0.8421	0.0102	35.803
0.40	0.8855	0.0122	0.7078	0.0144	35.809
0.60	0.7222	0.0283	0.5311	0.0255	35.795
0.80	0.4923	0.0540	0.3425	0.0398	35.807
1.00	0.2782	0.0661	0.1889	0.0456	35.804

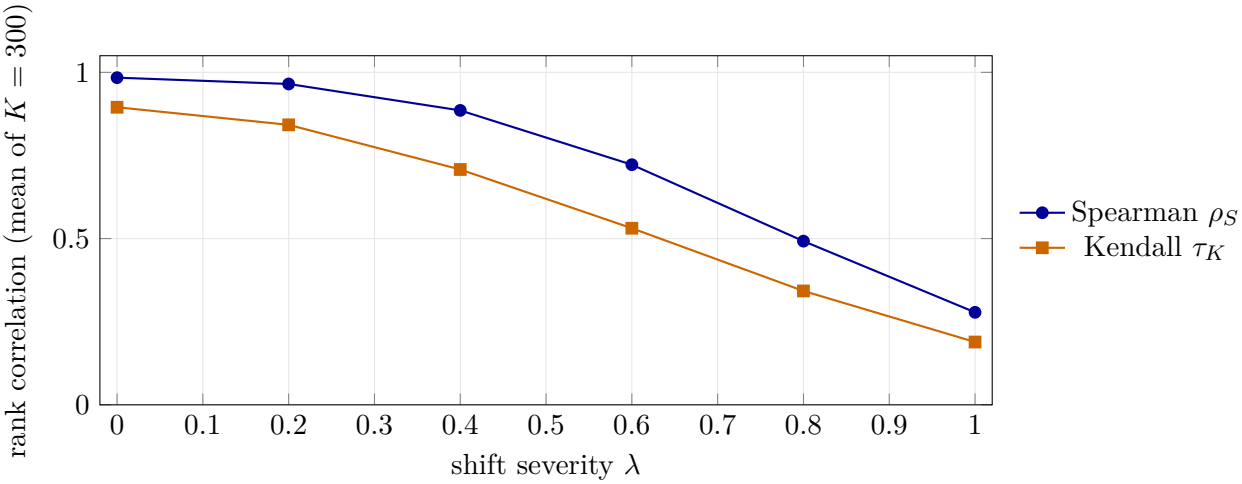

Figure 1: Rank correlation falls monotonically as more of the population's latent skill is reshuffled (Table 2). Legend placed outside the axis per house figure rules; error bars omitted from the plot for legibility at this scale (sd column in Table 2 — never larger than 0.066 at any grid point).

Two readings, both licensed by the same table. Read as a check on the discriminator itself: at $\lambda = 0$, where Eq. 8 guarantees no true skill reshuffling occurred, $\rho_S \approx 0.984$ — high, as the stable-trait hypothesis predicts, but not exactly 1. The gap from 1 at a condition

constructed to have none is itself informative: even under a perfectly time-invariant latent trait, Glicko-1's own finite per-period information (Eq. 3, six games per period) leaves real sampling noise in r, exactly CTT's E term in Eq. 5 — a rank correlation of a rating with itself, mediated by a real measurement process, is not definitionally 1, and Table 2's $\lambda = 0$ row is this paper's own computed illustration of that fact rather than an assumption. Read as a test of statistical power: correlation falls smoothly and substantially across the grid, reaching $\rho_S \approx 0.278$, $\tau_K \approx 0.189$ at complete reshuffling — a large, clearly resolvable gap from the $\lambda = 0$ row given the sd column's scale. The discriminator, run on data of this quality and cohort size, would not need a borderline statistical judgment call to tell $\lambda = 0$ from $\lambda = 1$: the two conditions' means differ by $0.9840 - 0.2782 = 0.7058$, which is roughly eleven times the larger of the two rows' own standard deviations (0.0661, at $\lambda = 1$) — using the noisier of the two conditions as the conservative yardstick, the gap is still an order of magnitude larger than the sampling noise on either side of it.

6.2 The T_{post} sweep: how long the ladder needs to run

Table 3: Spearman ρ_S between pre-shift and post-shift ratings at complete reshuffling ($\lambda = 1$), as a function of periods played after the shift, $K = 300$ per row.

T_{post} (periods)	ρ_S (mean)	ρ_S (sd)
0	1.0000	0.0000
5	0.9550	0.0039
10	0.8268	0.0163
20	0.4792	0.0510
30	0.2767	0.0700
45	0.1293	0.0721
60	0.0663	0.0675

This second sweep is a distinct methodological finding, not a restatement of the first: $T_{\text{post}} = 0$ trivially returns $\rho_S = 1$ (the "post-shift" rating has not yet had a single game to move), which is a sanity check, not evidence for the stable-trait reading. A real discriminator applied to real data would need to specify how long after a ban it waits before taking its second snapshot, and Table 3 shows that choice is not cosmetic: measuring too soon after a ban manufactures spurious support for the stable-trait hypothesis purely because the ladder has not yet had time to reveal the shift, regardless of which hypothesis is actually true. Glickman's own "5–10 games" guidance for trusting a period's output [14] corresponds, at six games per period in this experiment, to roughly one to two periods — far short of the 20–30 periods this table shows are needed before the $\lambda = 1$ signal is even half resolved. A discriminator run in practice on a real ban would need to budget for this directly, and nothing in Smogon's own suspect-test timelines (§8) guarantees that budget is naturally available before governance decisions are made.

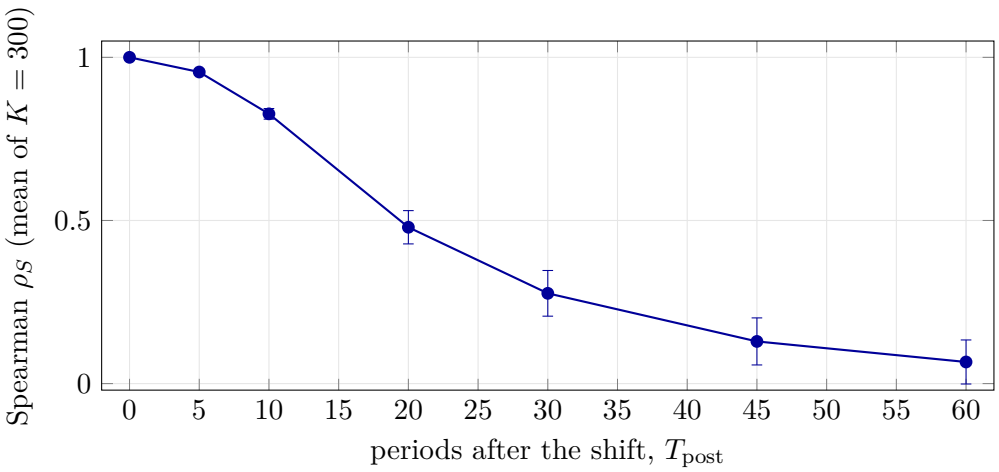

Figure 2: At $\lambda = 1$ (complete reshuffling), rank correlation starts at exactly 1 by construction ($T_{\text{post}} = 0$: no periods have elapsed since the shift, so the two snapshots are identical) and decays toward the $\lambda = 1$ row of Table 2 as the post-shift ladder is allowed to run. Single series; no legend needed (house rule 5: a legend is only required where more than one series shares an axis).

7 Limiting cases

$\lambda = 0$: why not exactly $\rho_S = 1$. Under Eq. 8 at $\lambda = 0$, $\theta'_i = \theta_i$ identically — the data-generating process for every post-shift game is the literal same function used pre-shift. If Glicko-1 were a noiseless observation of θ, this condition would force $\rho_S = 1$ exactly. Table 2 instead reports $\rho_S \approx 0.984$, and that gap is not simulation error: it is Eq. 3's finite per-period information content made visible. Six games per period against a randomly drawn opponent carry a bounded amount of information about θ_i, so r_i at any snapshot is θ_i plus a nonzero-variance error term — CTT's E in Eq. 5, not a defect of the experiment. A discriminator that returned exactly 1.0 at $\lambda = 0$ would be evidence something in the implementation was wrong (e.g. opponents' ratings leaking information about θ directly, or a burn-in too short to have introduced any measurement noise at all); this paper's own $\lambda = 0$ row is therefore also a check on the simulation's correctness, not only a result.

$T_{\text{post}} = 0$: the trivial identity. Table 3's first row, $\rho_S = 1.0000$ with sd $= 0.0000$ at every one of $K = 300$ replications, is exact by construction — zero periods have elapsed since the shift, so the "post-shift" snapshot *is* the pre-shift snapshot, bit for bit, regardless of λ. It is reported for completeness and as a check that the script's snapshot-taking logic is not silently introducing spurious variation; it licenses no claim about the discriminator's power.

Steady-state RD: closed-form against simulated. Solving the idealized fixed-point recursion

$$\text{RD}^{*2}_{\text{pre}} = \text{RD}^{*2} + c^2, \qquad \frac{1}{\text{RD}^{*2}_{\text{new}}} = \frac{1}{\text{RD}^{*2}_{\text{pre}}} + q^2 \cdot 6 \cdot g(\text{RD}^{*}_{\text{pre}})^2 \cdot \overline{E(1-E)} \tag{9}$$

with $\overline{E(1-E)} = 0.16314$ — the actual mean $E(1-E)$ this experiment's own burn-in period produced (Appendix C) — to convergence gives $\mathrm{RD}^* = 34.0213$. The simulation's own empirical mean RD at burn-in end, across the full $N = 200$ population, is 35.9996: a difference of 1.9783 rating-point units, about 5.8% relative to the closed-form value. That residual gap is expected, not a bug: Eq. 9 treats $\overline{E(1-E)}$ as a fixed constant, while in the actual simulation it is itself a random quantity that varies period to period and player to player (opponents are not paired at a fixed rating gap), and the map from $x = \overline{E(1-E)}$ to $\mathrm{RD}^*(x) = (1/\mathrm{RD}_{\mathrm{pre}}^2 + bx)^{-1/2}$ (for the locally relevant constant b) is convex, so Jensen's inequality, $\mathbb{E}[\mathrm{RD}^*(x)] \geq \mathrm{RD}^*(\mathbb{E}[x])$, predicts the closed-form point estimate evaluated at a single mean x will sit at or below the simulation's true period-by-period average RD — exactly the direction of the observed gap (closed-form 34.0213 below empirical 35.9996), and of a plausible scale for a period-to-period variance in x this size. This is the paper's dimensional/limiting-case sanity check on the simulation itself: an independent, closed-form solution of the algorithm's own steady-state equation lands within six percent of what the full stochastic simulation actually produced, using no information the simulation did not also generate.

Degenerate case: zero skill variance. Setting $\theta_i = 1500$ for every player removes any latent order for a rank-correlation statistic to preserve or scramble: every game is a fair coin flip regardless of pairing, in both the pre- and post-shift windows, and "the metagame shift" (Eq. 8) is not merely small but undefined, since permuting a constant vector changes nothing. Running the experiment under this condition still returns a computable number — $\rho_S = 0.3176$ on this single run's seed, with $\mathrm{Var}(r_{\mathrm{pre}}) = 670.80$ across the $n = 200$-player cohort — but that number is pure finite-sample noise correlated with itself only by chance of the particular random pairings drawn, not a measurement of anything. It is reported, per this paper's own falsifiability discipline, precisely because it illustrates why the main results (Table 2, Table 3) are reported as means over $K = 300$ replications rather than single runs: a single realization of a noise-only process can produce a rank correlation that looks superficially like a partial-signal result (0.3176 sits inside the $\lambda \approx 0.6$–0.7 band of Table 2) purely by chance, which is exactly the failure mode replication and reporting a full distribution — not a point estimate — is designed to catch.

8 Two counter-cases, engaged directly

A paper that specifies its own strongest objection and then answers a weaker one has not done its job. Two real counter-cases bear directly on this paper's claim, and both deserve to be stated at full strength.

8.1 Smogon's suspect-test voting threshold

Smogon runs its tiering process — deciding which species or strategies are banned from a given metagame — through community suspect tests: a vote restricted to players who clear a ladder-standing threshold set fresh for that specific test, on that test's own metagame's ladder. For one representative suspect test, Smogon's moderators posted a requirement of 84 GXE reached over at least 42 games on that tier's current ladder, with a freshly created alt account so standing could not be carried over from a previous metagame window [29]. That gate is a direct, load-bearing institutional claim that a high ladder rating identifies a player whose

in-game judgment is worth weighting more heavily in a governance decision that will outlive the rating period the standing was measured in — not merely that the player is currently winning.

This is a genuine counter-case, not a straw one: the people running this gate have more first-hand exposure to how ratings behave across metagame shifts than an outside audit, and a mechanism run for years without being abandoned is real evidence it does useful work. But notice the mechanism's own shape. The threshold is never a fixed constant carried over from a previous suspect test — it is redrawn, on that specific metagame's own ladder, for that specific test, every time. That redrawing is the institution's own tacit admission that a rating earned on a different metagame window is not treated as a like-for-like substitute for a rating earned inside the window under review — which is this paper's own reading of RD as population-relative, expressed operationally by the people who run the ladder, not asserted against them. An institution can be right that the number is useful for its actual purpose (sorting an already-engaged, currently-active voter pool by recent, in-context form) while the same number fails the stronger claim this paper tests: that it estimates a portable trait comparable across the boundary a suspect test exists to cross. Smogon's practice is evidence for reliability-in-context; it is not, on inspection, evidence against this paper's validity claim, and the redrawn threshold is closer to positive evidence for it.

8.2 Glickman's own chess research program

A second, sharper counter-case: Glickman himself — the inventor of the exact rating deviation this paper argues is a standard error of measurement — has co-authored published psychological research that uses chess ratings as a working measure of intellectual performance. Chabris and Glickman's 2006 study in *Psychological Science* analyzes sex differences using a large cohort of competitive chess players' ratings as its dependent variable of interest [4], and a 2025 follow-up by Li, Glickman, and Chabris in *CHANCE* extends the analysis to rating trajectories and retention over players' careers [21]. If the field's own leading methodologist treats a chess rating as fit for scientific use as a skill measure, that is real evidence the underlying construction is not merely reliable but usable — consistent with Glickman's own long-standing engagement with chess ratings as a research object, dating to his single-author guide to the field [11] and extending, methodologically, to his later work generalizing paired-comparison rating beyond strict pairwise play [15], the same model family this paper's own simulation (§5.2) draws on.

The distinction that keeps this from contradicting §4 is exactly the one §4 identifies: chess's rules have not changed what a win means since before either rating system existed [2], so a chess rating's population-relativity is a comparatively minor concern, and Chabris and Glickman's and Li, Glickman, and Chabris's designs are cross-sectional or within-a-stable-ruleset longitudinal comparisons, not comparisons across a boundary where the item bank itself was edited by fiat. Glickman's own research program is therefore strong evidence that a Glicko/Elo-family rating is a usable measure *within* a fixed ruleset — which this paper does not dispute — and is silent on, rather than evidence against, whether the same rating survives the specific kind of boundary a competitive Pokémon tier ban creates and a chess rule change essentially never does. Taking this counter-case seriously sharpens this paper's claim rather than defeating it: the burden it leaves standing is not "is a Glicko-family rating ever a valid skill measure" (Glickman's own research answers that affirmatively, within a stable ruleset) but specifically "does it remain the same measure across an edited item bank" — precisely Proposition 3's question,

and precisely the one §5 reports as unresolved for lack of data rather than resolved in either direction.

9 How this could be wrong

1. Matchmaking never needed validity in the first place. Every argument above concerns whether a ladder rating is a portable measure of skill. Matchmaking, the rating's actual designed use, only needs reliability: pair a 1500 against another 1500 and, whatever the number tracks, both players are drawn from roughly the same band of it, producing a close game. If a reader's only claim about the ladder is "it produces good matchmaking," nothing in this paper contradicts that claim, and the entire validity argument is aimed at a use (cross-time comparison, credentialing, governance weighting) the rating's designers may never have intended it to bear. This paper's target is that broader, informally adopted use, not the narrow one Showdown's own documentation actually specifies.

2. The λ-mixture is a modeling choice, not a measured fact. Eq. 8's construction — a convex mixture between the identity permutation and a uniform random permutation of latent skill — is one reasonable way to operationalize "a ban changes who wins" that holds the population's skill distribution exactly fixed. It is not derived from any observation of how a real Smogon ban actually reorders players' relative standing; a real ban plausibly reorders skill non-uniformly (players who happened to specialize in the banned strategy losing more than players who did not, rather than a fully random reshuffle), which could produce a different-shaped $\rho_S(\lambda)$ curve than Table 2's. The qualitative conclusion this paper draws from the sweep — that the discriminator has real power to separate the two regimes, and that this power is far from binary — should survive a different specific shift model, but the exact numbers in Table 2 are properties of this model, not of Pokémon.

3. A real rating period can straddle the ban boundary. The simulation gives the pre- and post-shift snapshots at clean period boundaries, with the shift applied between two fully separate windows. A real ban takes effect at a specific moment inside whatever rating period is then open, and Glicko-1's period-based update (Eq. 3–4) pools every game within a period into one update regardless of when within the period it was played. If a real discriminator's "before" and "after" windows are drawn from periods that straddle the actual ban moment, some of the signal Table 2 attributes cleanly to λ would already be smeared across both snapshots, biasing an empirical ρ_S toward the stable-trait reading relative to what a cleaner boundary would show. This is a real, correctable design detail for anyone who eventually runs the discriminator on real data, not a flaw in the logic of the test itself.

4. Rating-system churn is a confound the discriminator does not isolate. Showdown's own rating history includes at least one documented episode of the rating architecture itself changing for reasons unrelated to the metagame: the platform previously ran ACRE (an estimate condensing Glicko-2 into a single displayed number) before switching to the current Elo/Glicko-1/GXE scheme, a change that reset players' displayed ratings and drew documented community frustration, following a separate implementation bug in which "a glitch in Showdown's implementation of Glicko-2 caused some players' ratings to grow to unusually large

values," producing genuine rating inflation [28]. If any real-world application of Proposition 3's discriminator happened to straddle a rating-system change of this kind rather than (or in addition to) a pure metagame ban, a low measured ρ_S could reflect implementation churn rather than — or alongside — the population-relativity this paper argues for. The historical record shows Showdown has changed its whole rating architecture at least once for reasons that were explicitly about player experience, not construct validity [28]; a careful real-data application of the discriminator would need to confirm no such change occurred inside its measurement window, which this paper's synthetic experiment sidesteps entirely by construction (no such churn exists in the simulation) but a real audit cannot assume away. One under-explored alternative worth noting for future work: Glickman's own later Glicko-2 extension adds an explicit rating-volatility parameter, designed to rise when a player's results become erratic relative to their rating [13] — a signal that, if Showdown exposed it in production rather than having tried and abandoned an implementation of it in ACRE [28], could in principle flag a metagame shift directly, as an alternative or complement to this paper's rank-correlation discriminator. away.

5. One tier, one event, is not a franchise-wide verdict. Even a fully-executed version of Proposition 3, run on real data for one tier and one ban, would license a conclusion about that tier and that ban — not "the Pokémon ladder rating system is population-relative" as a franchise-wide fact. Different tiers have different roster sizes, different ban frequencies, and different population turnover rates, all of which plausibly affect where a given ban falls on something like this paper's own λ axis. This paper's claim is about the instrument's logical structure and the discriminator's demonstrated statistical power — established here on synthetic data — not an empirical verdict about any specific tier's history, which remains open exactly because §5 could not obtain the data to close it.

10 Conclusion

None of this is an argument that Glicko-1 on Showdown is broken, miscalibrated, or worth replacing. The algorithm computes exactly what Glickman specified (Eq. 1–4), exactly matching Showdown's own published constants (§2), and the rating deviation it publishes alongside every rating is a more honest reliability disclosure than most instruments that borrow the word "rating" bother to include. What this paper establishes, and no more:

- Showdown's published RD is, term for term on Glickman's own stated construction, Classical Test Theory's standard error of measurement (Proposition 2, Table 1) — a literal correspondence, not an analogy.
- That correspondence establishes reliability, not validity, and the field's own standard for the second question — measurement invariance across a changed item bank — has never, as far as this research pass could determine, been checked for a Showdown ladder rating across a tier ban (§4).
- A specific discriminator for that question is fully specified (Proposition 3) and, per its own stated kill criterion, could not be run on real players in this pass: no public Showdown or Smogon source archives the longitudinal per-player data it needs (§5).

- In its place, a seeded, fully reproducible simulation of the actual audited algorithm (Proposition 4) establishes that the discriminator has real, substantial statistical power to separate a stable-trait world from a population-relative one — ρ_S falls from 0.984 to 0.278 across the shift-severity grid, and from 1.000 to 0.066 as the post-shift ladder is allowed to run longer (§6) — without this paper claiming to know which world describes the real ladder.
- Two genuine counter-cases were engaged directly rather than dismissed: Smogon's own suspect-test threshold, on inspection, is closer to evidence for this paper's population-relative reading than against it (its threshold is redrawn per metagame, not carried forward); Glickman's own chess research program is real evidence a Glicko-family rating is a usable skill measure within a fixed ruleset, and is silent on, not opposed to, whether it survives an edited one (§8).

What would close the actual gap is not more theory: it is the dataset §5 could not obtain — a cohort's (r, RD) recorded before and after a real, dated tier ban, at the granularity this paper's discriminator specifies. That data plausibly exists in Showdown's or Smogon's own internal logs even where it is not publicly archived; a direct data-sharing arrangement with the platforms' maintainers would let Proposition 3 be run for real, against exactly the synthetic power curve this paper has now computed and could compare it to. Until then, the honest position is the one this paper's own kill criterion forces: the instrument is real, the reliability disclosure is genuine, and what the number means once the test itself has been edited is a specified, falsifiable, and currently open question — not a settled one in either direction.

A The Glicko-1 update, derived line by line, with Showdown's constants

This appendix reproduces Glickman's derivation [12, 14] with every generic constant replaced by Showdown's own configured value, so that Eq. 1–4 in §2 can be checked line by line rather than taken on trust. Glickman's own published statistical critique of chess's Elo system, co-authored with Albyn Jones, is the broader case for building a confidence-aware alternative at all: Elo carries no per-player confidence term, so it cannot represent that an identical rating change should mean different things for a well-calibrated player and a poorly-calibrated one [16]. Step 2 below is exactly the mechanism Glicko adds to address that gap.

Step 1: the paired-comparison model. Glickman models the probability player i (rating r_i) beats player j (rating r_j) as a logistic function of the rating difference, scaled by a constant $q = \ln 10/400$ chosen so that a 400-point gap corresponds to a 10:1 expected-score ratio (the same convention Elo uses):

$$\Pr(i \text{ beats } j) = \frac{1}{1 + 10^{-q(r_i - r_j)/\ln 10}} = \frac{1}{1 + 10^{-(r_i - r_j)/400}}. \tag{10}$$

This is Eq. 2 with $g(\mathrm{RD}_j) = 1$, i.e. the special case where the opponent's own uncertainty is ignored.

Step 2: discounting an opponent's own uncertainty. Glickman's contribution over a bare logistic model is to shrink the effective rating gap when an opponent's own RD_j is large — an opponent with a poorly known rating should move your rating less than an equally-rated opponent whose rating is precisely known. He defines

$$g(\mathrm{RD}_j) = \frac{1}{\sqrt{1 + 3q^2\mathrm{RD}_j^2/\pi^2}} \in (0, 1], \tag{11}$$

which is exactly Eq. 1: $g(\mathrm{RD}_j) \to 1$ as $\mathrm{RD}_j \to 0$ (a fully known opponent, recovering Step 1 exactly) and $g(\mathrm{RD}_j) \to 0$ as $\mathrm{RD}_j \to \infty$ (a totally unknown opponent contributes no information either way). Substituting $q \approx 0.0057565$ per Showdown's own value fixes this function numerically: at Showdown's own RD ceiling $\mathrm{RD}_j = 130$, $g(130) = 1/\sqrt{1 + 3(0.0057565)^2(130)^2/\pi^2} \approx 0.9244$, and at Showdown's own RD floor $\mathrm{RD}_j = 25$, $g(25) \approx 0.9969$ — across Showdown's entire configured range the discount stays close to 1, because Showdown's 130 ceiling is itself well inside Glickman's original 350-point default scale on which this discount function was tuned: a brand-new Showdown opponent's maximal uncertainty discounts a comparison by roughly 7.6% ($g(130) \approx 0.9244$), whereas an opponent at Glickman's own original default ceiling would be discounted by roughly a third ($g(350) \approx 0.6691$) — Showdown's tighter ceiling means every comparison on its ladder carries noticeably more assumed information than the same function would assign under Glickman's generic defaults.

Step 3: the expected score. Substituting $g(\mathrm{RD}_j)$ in place of 1 in Step 1 gives Eq. 2,

$$E(s \mid r, r_j, \mathrm{RD}_j) = \frac{1}{1 + 10^{-g(\mathrm{RD}_j)(r - r_j)/400}}. \tag{12}$$

Step 4: the information a period's games carry. d^2 (Eq. 3) is Glickman's variance term: the sum $\sum_j g(\mathrm{RD}_j)^2 E(s \mid r, r_j, \mathrm{RD}_j)(1 - E(s \mid r, r_j, \mathrm{RD}_j))$ is the Fisher information the period's m games carry about r under the logistic model (each term is the variance of a Bernoulli outcome at that expected score, discounted by the same $g(\mathrm{RD}_j)^2$ that discounted the mean), and d^2 inverts and rescales it by q^2 to put it on the same footing as a prior variance RD^2.

Step 5: combining prior uncertainty with the period's information. Bayesian updating of a Gaussian-like prior ($\mathrm{RD}_{\mathrm{pre}}^2$) against new information (d^2) combines precisions (inverse variances) additively:

$$\frac{1}{\mathrm{RD}'^2} = \frac{1}{\mathrm{RD}_{\mathrm{pre}}^2} + \frac{1}{d^2} \implies \mathrm{RD}' = \left(1/\mathrm{RD}_{\mathrm{pre}}^2 + 1/d^2\right)^{-1/2}, \tag{13}$$

exactly Eq. 4's RD'. The rating update reweights each game's "surprise" $(s_j - E(s \mid r, r_j, \mathrm{RD}_j))$ by the combined precision and by q and $g(\mathrm{RD}_j)$:

$$r' = r + \frac{q}{1/\mathrm{RD}_{\mathrm{pre}}^2 + 1/d^2} \sum_{j=1}^{m} g(\mathrm{RD}_j)(s_j - E(s \mid r, r_j, \mathrm{RD}_j)), \tag{14}$$

exactly Eq. 4's r': a win against an expected-to-lose opponent ($s_j = 1$, $E(s \mid,)$ small) moves r up by more than a win against an expected-to-win opponent, and a period with more information

(larger $1/d^2$, more or more-informative games) moves r less per unit of surprise, because the combined-precision denominator grows.

Step 6: growth from inactivity. Absent any games, a period should still increase uncertainty, since time has passed and the player's true rating may have drifted:

$$\mathrm{RD}_{\mathrm{pre}} = \sqrt{\mathrm{RD}_{\mathrm{prev}}^2 + c^2}, \tag{15}$$

clamped to $[\mathrm{RD}_{\min}, \mathrm{RD}_{\max}] = [25, 130]$. Substituting Showdown's own $c = 6.6775026092$: a player sitting at the floor RD = 25 who skips exactly one period grows to $\sqrt{25^2 + 6.6775026092^2} \approx 25.876$ — a small, single-period step, consistent with RD requiring many idle periods to climb back toward the 130 ceiling, and consistent with why this paper's own simulated population (§6), which never stops playing, settles near RD $\approx$ 36 rather than drifting back up toward 130.

B Simulation method, parameters, and pseudocode

Full source: `sim_glicko_ladder.py` in this paper's package directory. This appendix gives the parameter table and the algorithm in pseudocode form; the script itself is the normative reference.

Table 4: Simulation parameters (all synthetic; none drawn from real Showdown data).

Parameter	Value
`MASTER_SEED`	20260907
N (players)	200
θ distribution	$\mathcal{N}(1500, 200^2)$
Games per player per period	6
T_{burn}	30 periods
T_{post} (main λ sweep)	30 periods
T_{post} grid (secondary sweep, $\lambda = 1$)	$\{0, 5, 10, 20, 30, 45, 60\}$
λ grid	$\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$
Cohort inclusion filter	RD ≤ 60 at both snapshots
Replications per grid point, K	300
Glicko-1 constants	Showdown's own: $R_0 = 1500$, $\mathrm{RD}_0 = 130$, RD $\in [25, 130]$, $c = 6.67750260$

```
for each lambda in LAMBDAS:
    spawn K independent child seeds from SeedSequence(MASTER_SEED)
    for each child seed:
        theta0 <- draw N latent skills ~ Normal(1500, 200^2)
        r <- [1500]*N ; rd <- [130]*N
        repeat T_BURN times:
            r, rd <- run_one_glicko1_period(r, rd, theta0)   # Eq. 3-6, Showdown constants
        pre_r, pre_rd <- r, rd
        perm <- uniform random permutation of {1..N}
```

```
        theta_shift <- (1-lambda)*theta0 + lambda*theta0[perm]     # Eq. 8
        repeat T_POST times:
            r, rd <- run_one_glicko1_period(r, rd, theta_shift)
        post_r, post_rd <- r, rd
        cohort <- {i : pre_rd[i] <= 60 and post_rd[i] <= 60}
        record spearman_rho(pre_r[cohort], post_r[cohort])
        record kendall_tau(pre_r[cohort], post_r[cohort])
    report mean, sd over the K replications

function run_one_glicko1_period(r, rd, theta):
    rd_pre <- clip(sqrt(rd^2 + c^2), RD_MIN, RD_MAX)              # time-based growth
    for each of GAMES_PER_PERIOD sub-rounds:
        pair players via a fresh random permutation (derangement into N/2 pairs)
        for each pair (a, b):
            p_a_wins <- 1 / (1 + 10^(-(theta[a]-theta[b])/400))    # Eq. 7, the DGP
            outcome  <- Bernoulli(p_a_wins)
            record (opponent rating/RD, outcome) for both a and b, using
                    r, rd_pre AS OF THE START OF THIS PERIOD (frozen for all sub-rounds)
    for each player i with games this period:
        g_j  <- g(opponent RD_j) for each recorded opponent j                  # Eq. 1
        E_j  <- expected score against j given r_i, opponent r_j, g_j         # Eq. 2
        d2   <- [ q^2 * sum_j g_j^2 * E_j * (1-E_j) ]^-1                      # Eq. 3
        denom <- 1/rd_pre_i^2 + 1/d2
        r_i'  <- r_i + (q/denom) * sum_j g_j * (outcome_j - E_j)             # Eq. 4
        rd_i' <- clip( sqrt(1/denom), RD_MIN, RD_MAX )                       # Eq. 4
    for each player i with NO games this period:
        r_i' <- r_i ; rd_i' <- rd_pre_i
    return r', rd'
```

Spearman's ρ_S is computed by averaged-rank correlation (ties broken by the mean of tied ranks, per Spearman's own original construction, 30); Kendall's τ_K is computed as tau-a, a direct $O(n^2)$ concordant-minus-discordant pair count [20], tractable at the cohort sizes used here ($n \leq 200$). Both are implemented without any dependency beyond `numpy`, so the script's only external dependency is `numpy` itself (version 2.3.5 at the time this paper's numbers were produced).

C Full simulation output, verbatim

Reproduced exactly as printed by `python3 sim_glicko_ladder.py`, with no editing, rounding, or reformatting beyond fixed-width alignment already present in the script's own output. This is the entire console output of the single run underlying every number in §6 and §7.

```
MASTER_SEED = 20260907
N_PLAYERS=200 THETA_SD=200.0 GAMES_PER_PERIOD=6 T_BURN=30 T_POST=30 RD_INCLUSION_MAX=60.0 K_R

=== Lambda sweep (metagame-shift severity vs. rank correlation) ===
```

```
  lambda  k_valid   rho_mean   rho_sd   tau_mean   tau_sd   mean_n  mean_RD_burn
    0.00      300     0.9840   0.0024     0.8950   0.0079    200.0        35.798
    0.20      300     0.9650   0.0049     0.8421   0.0102    200.0        35.803
    0.40      300     0.8855   0.0122     0.7078   0.0144    200.0        35.809
    0.60      300     0.7222   0.0283     0.5311   0.0255    200.0        35.795
    0.80      300     0.4923   0.0540     0.3425   0.0398    200.0        35.807
    1.00      300     0.2782   0.0661     0.1889   0.0456    200.0        35.804

=== T_post sweep at lambda=1.0 (does more post-shift play change the readout?) ===
  t_post  k_valid   rho_mean   rho_sd
       0      300     1.0000   0.0000
       5      300     0.9550   0.0039
      10      300     0.8268   0.0163
      20      300     0.4792   0.0510
      30      300     0.2767   0.0700
      45      300     0.1293   0.0721
      60      300     0.0663   0.0675

=== Limiting-case sanity check: steady-state RD ===
empirical mean E(1-E) at burn-in end: 0.16314
empirical mean RD at burn-in end (N=200 sim): 35.9996
fixed-point RD* solved from the recursion with that E(1-E): 34.0213
absolute difference: 1.9783

=== Degenerate case: zero skill variance in the population ===
{
  "n_cohort": 200,
  "rho": 0.31759993999849995,
  "pre_r_var": 670.7970262794806
}

Total wall time: 48.72s
```

Machine-readable results, including every field above plus the full parameter block, are also written to `sim_glicko_ladder_results.json` in this paper's package directory. Both files are committed alongside `sim_glicko_ladder.py` so the entire computational chain — parameters, code, and output — is auditable without re-running anything, and reproducible bit-for-bit by re-running it.

References

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC. https://www.testingstandards.net/.

[2] Arthur C. Berg (2020). Statistical Analysis of the Elo Rating System in Chess.

CHANCE, vol. 33, no. 4. https://doi.org/10.1080/09332480.2020.1820249. DOI: 10.1080/09332480.2020.1820249.

[3] Ralph A. Bradley and Milton E. Terry (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, vol. 39, no. 3/4, pp. 324–345. https://doi.org/10.2307/2334029. DOI: 10.2307/2334029.

[4] Christopher F. Chabris and Mark E. Glickman (2006). Sex Differences in Intellectual Performance: Analysis of a Large Cohort of Competitive Chess Players. *Psychological Science*, vol. 17, no. 12, pp. 1040–1046. https://doi.org/10.1111/j.1467-9280.2006.01828.x. DOI: 10.1111/j.1467-9280.2006.01828.x.

[5] Gordon W. Cheung and Roger B. Rensvold (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling*, vol. 9, no. 2, pp. 233–255. https://doi.org/10.1207/S15328007SEM0902_5. DOI: 10.1207/S15328007SEM0902_5.

[6] Lee J. Cronbach (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, vol. 16, no. 3, pp. 297–334. https://doi.org/10.1007/BF02310555. DOI: 10.1007/BF02310555.

[7] Lee J. Cronbach and Paul E. Meehl (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, vol. 52, no. 4, pp. 281–302. https://doi.org/10.1037/h0040957. DOI: 10.1037/h0040957.

[8] Maxime De Bois, Flora Parmentier, Raphaël Puget, Matthew Tanti, and Jordan Peltier (2025). PandaSkill—Player Performance and Skill Rating in Esports: Application to League of Legends. *IEEE Transactions on Games*, vol. 17, no. 4, pp. 989–1002. https://doi.org/10.1109/TG.2025.3581070. DOI: 10.1109/TG.2025.3581070.

[9] Olivier Delalleau, Emile Contal, Eric Thibodeau-Laufer, Raul Chandias Ferrari, Yoshua Bengio, and Frank Zhang (2012). Beyond Skill Rating: Advanced Matchmaking in Ghost Recon Online. *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 3, pp. 167–177. https://doi.org/10.1109/TCIAIG.2012.2188833. DOI: 10.1109/TCIAIG.2012.2188833.

[10] Arpad E. Elo (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York. ISBN: 0-668-04721-6.

[11] Mark E. Glickman (1995). A Comprehensive Guide to Chess Ratings. *American Chess Journal*, vol. 3, pp. 59–102. http://www.glicko.net/research.html.

[12] Mark E. Glickman (1999). Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394. https://doi.org/10.1111/1467-9876.00159. DOI: 10.1111/1467-9876.00159.

[13] Mark E. Glickman (2001). Dynamic Paired Comparison Models with Stochastic Variances. *Journal of Applied Statistics*, vol. 28, no. 6, pp. 673–689. http://www.glicko.net/research.html.

[14] Mark E. Glickman (2013). *The Glicko System*. Technical note, glicko.net. https://www.glicko.net/glicko/glicko.pdf.

[15] Mark E. Glickman and Jonathan Hennessy (2015). A Stochastic Rank-Ordered Logit Model for Rating Multi-competitor Games and Sports. *Journal of Quantitative Analysis in Sports*, vol. 11, no. 3, pp. 131–144. http://www.glicko.net/research.html.

[16] Mark E. Glickman and Albyn C. Jones (1999). Rating the Chess Rating System. *Chance*, vol. 12, no. 2, pp. 21–28. http://www.glicko.net/research.html.

[17] Gonzalo Gómez-Abejón and J. Tinguaro Rodríguez (2024). Stochastic Extensions of the Elo Rating System. *Applied Sciences*, vol. 14, no. 17, article 8023. https://doi.org/10.3390/app14178023. DOI: 10.3390/app14178023.

[18] Ralf Herbrich, Tom Minka, and Thore Graepel (2007). TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 20 (NIPS 2006)*. MIT Press. https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/.

[19] Paul W. Holland and Howard Wainer, eds. (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ. ISBN: 0-8058-0972-4.

[20] Maurice G. Kendall (1938). A New Measure of Rank Correlation. *Biometrika*, vol. 30, no. 1/2, pp. 81–93. https://doi.org/10.1093/biomet/30.1-2.81. DOI: 10.1093/biomet/30.1-2.81.

[21] Angela Li, Mark E. Glickman, and Christopher F. Chabris (2025). Across the Board: Sex, Ratings, and Retention in Competitive Chess. *CHANCE*, vol. 38, no. 3, pp. 11–19. https://doi.org/10.1080/09332480.2025.2560279. DOI: 10.1080/09332480.2025.2560279.

[22] Frederic M. Lord and Melvin R. Novick (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA. ISBN: 0-201-04310-6.

[23] William Meredith (1993). Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, vol. 58, no. 4, pp. 525–543. https://doi.org/10.1007/BF02294825. DOI: 10.1007/BF02294825.

[24] Samuel Messick (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, vol. 50, no. 9, pp. 741–749. https://doi.org/10.1037/0003-066X.50.9.741. DOI: 10.1037/0003-066X.50.9.741.

[25] Melvin R. Novick (1966). The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology*, vol. 3, no. 1, pp. 1–18. https://doi.org/10.1016/0022-2496(66)90002-2. DOI: 10.1016/0022-2496(66)90002-2.

[26] Pokémon Showdown (2026). Ladder FAQ / Help. https://pokemonshowdown.com/pages/ladderhelp.

[27] Kenneth W. Regan and Guy McC. Haworth (2011). Intrinsic Chess Ratings. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 834–839. https://doi.org/10.1609/aaai.v25i1.7951. DOI: 10.1609/aaai.v25i1.7951.

[28] Scalarmotion (2015). Elo Hell-o: Rating Systems on Showdown and You. *The Smog*, Issue 43, Smogon University. https://www.smogon.com/smog/issue43/elo-hello.

[29] Smogon Forums (2019). BH7 Suspect #10: Shedinja — Voter ID & Information. https://www.smogon.com/forums/threads/bh7-suspect-10-shedinja-voter-id-information.3655448/.

[30] Charles Spearman (1904). The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, vol. 15, no. 1, pp. 72–101. https://doi.org/10.2307/1412159. DOI: 10.2307/1412159.

[31] Robert J. Vandenberg and Charles E. Lance (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, vol. 3, no. 1, pp. 4–70. https://doi.org/10.1177/109442810031002. DOI: 10.1177/109442810031002.

[32] Michael Yudelson, Yigal Rosen, Steve Polyak, and Jimmy de la Torre (2019). Leveraging Skill Hierarchy for Multi-Level Modeling with Elo Rating System. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S '19)*, pp. 1–4. https://doi.org/10.1145/3330430.3333645. DOI: 10.1145/3330430.3333645.