HEDONIX RESEARCH

WORKING PAPER · No. 2026-02

Extending the H6 Hedonic Engine: Pop-Count Integration, a Multi-Tier Model Ecosystem, and Out-of-Sample Validation

Evidence from the Scarlet & Violet, Sword & Shield, and Sun & Moon Eras (2017–2026)

The methodology described in this paper underlies the production H6 Hedonic Engine, the pricing core of Hedonix.

Philipp Baro Hedonix Research, Frankfurt am Main philipp@hedonix.tech · hedonix.tech

First version: May 2026 · This version: May 2026 (updated 2026-05-18 with §8.7 stale-price-bias addendum)

This paper updates and extends Hedonix Working Paper No. 2026-01 (Baro, 2026). The 2026-01 specification remains the documentary baseline. Results reported here supersede it where they overlap.

The 2026-05-18 revision adds §8.7 documenting a post-publication diagnostic finding: the Sharpe ratios reported in §5.5 are inflated by a stale-price / smoothing bias of approximately 1.5–2× relative to a monthly-return-based recomputation. The qualitative claim that Q5 has approximately twice the Sharpe of Q1 survives the correction; the absolute Sharpe levels do not.

Executive Summary

For readers without time for the full paper. The five takeaways:

1. **Pop-count integration closes the original residual gap.** Adding two PSA Population Report regressors (log total graded, gem rate) lifts in-sample R^2 from 0.879 to 0.914 and leave-one-set-out cross-validation R^2 from 0.79 to 0.87.
2. **A parallel PSA-9 model achieves comparable fit** (in-sample R^2 0.920, n=347), enabling cross-grade fair-value comparison and surfacing artist heterogeneity that differs across grade tiers.
3. **A non-parametric raw-price model on 2,622 cards** reaches honest LOSO R^2 0.83 with median absolute error 34%, using revealed-preference features (graded

population volumes, eBay sales velocity, sealed-product premium) but no PSA price information.

4. **The single-anchor live performance result of +60% return spread does not survive multi-anchor stress.** Across four monthly anchors (February to May 2025, 365-day forward windows), median Q5-Q1 spread is +3.94 percentage points and inverts directionally between earlier and later anchors. The single-anchor 2025-05-08 result that anchored prior marketing is the worst data point in the four-anchor sweep.

5. **The cross-sectional risk spread is robust at the 100% of-anchors level.** Q5 cohorts deliver median Sharpe approximately twice that of Q1 (2.88 vs. 1.60) and median annualized volatility approximately 40% lower (24% vs. 40%). The framework is defensible as a cross-sectional risk-management tool. It is not defensible as a directional alpha signal at the cohort level once anchor-position robustness is required. Sub-segment analysis indicates the cohort logic generates a robust return premium only in older-era cards, older-style rarities, and sub-$50 price points, the complement of the segment on which the current product surface is centred.

Plain-Language Summary

For readers who collect cards but don't read finance papers.

We built a model that scores Pokémon TCG cards into five risk buckets each day. The lowest-risk bucket has historically traded with about 40% less price volatility than the highest-risk bucket, and a much better return-per-unit-of-risk ratio (Sharpe ratio of 2.9 versus 1.6). This held in every test window we ran.

What we also tested: whether the lowest-risk bucket has higher *total returns* than the highest-risk bucket. Our first backtest said yes, by about 17 percentage points over a year. But when we re-ran the same test from three other starting points, the result didn't hold up. At later starting points the result actually inverted. So we don't tell you the lowest-risk bucket will outperform in price. We tell you it will swing less violently. That's a useful tool for risk management, not a tool for picking winners.

The full methodology, all the test details, and every negative finding are documented in the sections below.

Abstract

This paper updates and extends the hedonic pricing framework introduced in Hedonix Working Paper No. 2026-01 along five dimensions. First, we close the principal residual structure identified in the original paper by integrating PSA Population Report data into the production specification. Two new regressors — the log of total graded population and the PSA-10 gem rate — raise in-sample R^2 from 0.879 to 0.914 on a temporally aligned

snapshot of n=343 cards, and lift leave-one-set-out cross-validation R^2 from 0.79 to 0.87. The two regressors carry opposite signs and tell economically distinct stories: population is a liquidity proxy dominated by demand, while gem rate is a supply proxy dominated by production-side condition variance. Second, we report a parallel hedonic specification for PSA-9 prices on the same panel (n=347, in-sample R^2 0.920, LOSO R^2 0.745), allowing cross-grade fair-value comparison. Third, we expand the analytic universe from 360 cards to 2,635 cards spanning fifty-six expansion sets across three modern eras via a bulk historical pull of two-year transaction histories. On this expanded universe we develop a non-parametric raw-price model (XGBoost, n=2,622, LOSO R^2 0.83, median absolute percent error 34%) that uses revealed-preference features — graded population volumes, eBay daily sales volume, and sealed-product premia — but no PSA price information. Fourth, we report out-of-sample validation along four complementary axes: five-fold random cross-validation (R^2 0.839, median 17.5%), leave-one-set-out cross-validation (R^2 0.786, median 21.5%), a 180-day convergence backtest, and a single-anchor 365-day equal-weighted top-quintile portfolio reference (mean total return +60% vs. broad-universe mean +43%, spread +17 percentage points). Fifth, we run a post-hoc multi-anchor robustness sprint that re-fits the production model at four independent anchor dates and finds that the cross-sectional return spread is *not* multi-anchor robust: median Q5-Q1 spread across four anchors is +3.94 percentage points, and the spread inverts directionally between earlier and later anchors. The same sprint finds the cross-sectional *risk* spread is robust at the 100% of-anchors level: Q5 cohorts deliver median Sharpe approximately twice that of Q1 (2.88 vs. 1.60) and median annualized volatility approximately 40% lower (24% vs. 40%), in every anchor tested. We accordingly re-position the framework: defensible as a cross-sectional risk-management tool — the cohort signal robustly identifies lower-volatility, higher-Sharpe buckets — but *not* defensible as a directional alpha signal at the cohort level once anchor-position robustness is required. We discuss methodological implications, including the null effects of two stated-preference features (LLM-based artwork ratings and Google Trends interest), the structural plafond on raw-price fair-value modeling without a revealed-preference demand instrument, and a sub-segment analysis suggesting the cohort logic generates a robust return premium only in older-era cards, older-style rarities, and sub-$50 price points — the *complement* of the segment on which the current product surface is centred.

Keywords: hedonic pricing, collectibles, alternative assets, trading cards, cross-sectional asset pricing, PSA population, gradient boosting, walk-forward validation, Pokémon TCG

JEL Classification: G11, G12, G14, C52, Z11

---

1. Introduction

Hedonix Working Paper No. 2026-01 (Baro, 2026) reported a hedonic regression of the natural logarithm of PSA-10 sale prices on a curated set of structural and empirically constructed features, achieving $R^2 = 0.879$ on a sample of 360 Special Illustration Rare and Illustration Rare cards from fourteen Scarlet & Violet expansion sets. Section 6 of that

paper noted three residual concerns: the omission of any direct measure of supply scarcity, the limitation to a single product era, and the absence of an out-of-sample validation framework defensible to external readers. The present paper addresses each of these concerns.

The contribution is fourfold. First, we operationalize a custom PSA Population Report scraper that resolves card identities against the PSA specification catalog and pulls live population counts. The resulting two-feature extension — log total graded volume and PSA-10 gem rate — measurably improves in-sample fit (R^2 0.879 → 0.914) and, more importantly, lifts leave-one-set-out cross-validation R^2 from 0.79 to 0.87. The economic interpretation is non-trivial: in a cross-section of post-launch SV cards, the population variable carries the opposite sign from the textbook supply prediction. We document this finding and argue it reflects the endogeneity of grading-submission behavior with respect to demand.

Second, we extend the hedonic specification to a parallel PSA-9 target on the same sample. The PSA-9 model achieves in-sample R^2 0.920 and reproduces the broader pattern of significant character, artist, rarity, and pop-count effects with magnitudes that differ informatively across grade tiers.

Third, we substantially expand the analytic universe. A two-year bulk historical pull from a commercial pricing API yields per-card transaction histories for 2,619 of 2,635 targeted cards across fifty-six expansion sets in the Sun & Moon (2017–2019), Sword & Shield (2020–2023), and Scarlet & Violet (2023–2026) eras. The expanded sample makes feasible an entirely separate raw-price model trained as a gradient-boosted tree ensemble, which we use as an internal cross-grade consistency scanner.

Fourth, we document out-of-sample validation along four complementary axes: random k-fold cross-validation, leave-one-set-out cross-validation, a 180-day convergence backtest from October 2025 to April 2026, and the first 365 days of a live equal-weighted quintile portfolio. The convergence and live-portfolio results both support a specific interpretation: the framework provides defensible cross-sectional fair-value rankings useful for risk management and portfolio construction at the cohort level, but does not provide card-level directional forecasts.

The paper proceeds as follows. Section 2 describes the data updates. Section 3 introduces the H6 v2 specification with pop-count regressors. Section 4 reports the parallel PSA-9 specification. Section 5 documents the out-of-sample validation framework. Section 6 develops the raw-price XGBoost extension and its methodological trail through several null findings. Section 7 reports the live Hedonix Index results, including an anachronism-controlled re-fit. Section 8 discusses limitations and open research questions. Section 9 concludes.

2. Data Expansion

2.1 Universe scope

The 2026-01 analytic sample comprised 360 Special Illustration Rare and Illustration Rare cards from fourteen Scarlet & Violet expansion sets. The present paper retains that panel as the *hedonic core* and supplements it with a substantially expanded *research universe* of 2,635 cards spanning fifty-six expansion sets across three modern card-frame eras.

Table 1. Research-universe scope by era.

Era	Years	Sets	Cards
Sun & Moon	2017–2019	18	752
Sword & Shield	2020–2023	23	1,002
Scarlet & Violet	2023–2026	16	881
Total		**56**	**2,635**

The universe is restricted to investment-relevant rarity tiers: Special Illustration Rare, Illustration Rare, Hyper Rare, and Ultra Rare in the SV era; Rare Secret, Rare Rainbow, Rare Holo VMAX, Rare Holo VSTAR, Trainer Gallery, Radiant Rare, and Amazing Rare in SWSH; and Rare Holo GX, Rare Ultra, and inherited shiny/secret classes in S&M. Common, uncommon, standard rare, reverse-holo, energy, and basic trainer non-art versions are excluded.

2.2 PSA Population Report integration

To address the supply-scarcity gap identified in 2026-01 §6, we developed an in-house PSA Pop Report scraper. The pipeline operates in two stages: (i) a card-identity resolution stage that queries the PSA SpecSearch endpoint with set name and card number, caches the resolved PSA specification identifier per card, and applies a foreign-language-variant filter to exclude Indonesian, Thai, and Japanese alternate specs from the resolution; and (ii) a population-retrieval stage that pulls the live population JSON for each cached specification and writes counts to a dedicated table.

Population coverage is complete for the hedonic core (358 of 358 cards) and partial but extensive for the broader research universe.

For each card we record:

- `psa_total_graded`: total population across all PSA grades.
- `psa10_count`, `psa9_count`, `psa8_count`: counts at the three grades most economically relevant.
- `psa_gem_rate`: ratio of PSA-10 to total graded.

The scraper runs on a weekly cadence and accumulates a population time series for future use; in the cross-sectional analyses reported here, we use only the most recent snapshot.

2.3 Historical price warehouse

In addition to the live cross-section used in 2026-01, we constructed a historical price warehouse via a bulk pull from a commercial pricing aggregator using a 730-day history window. The pull yielded 578,919 daily price observations across 2,619 cards (99.4% coverage of the targeted universe) together with 2,448 per-card aggregate graded-sales snapshots covering PSA-10, PSA-9, PSA-8, CGC-10, CGC-8, BGS-10, and SGC-10 prices and volumes. The data are staged in a dedicated `historical` schema and feed the convergence backtest of Section 5.3, the live Hedonix Index of Section 7, and the raw-price model of Section 6.

A small number of mid-Sun-&-Moon cards (13 of 2,635) are absent from the commercial aggregator despite being present in the public card metadata API. These omissions are concentrated in the earliest sub-sets of the era and do not overlap with the hedonic core panel. We treat them as truly absent from the source rather than as matching errors.

---

3. H6 v2: Closing the Pop-Count Gap

3.1 Motivation: the aesthetic-quality residual revisited

Section 6.2 of Baro (2026) identified the model's incomplete measurement of aesthetic and scarcity-driven demand as the principal limitation of the original specification. The author conjectured that high-pop-count cards might appear undervalued to the original H6 because the model had no direct measure of grading supply.

The PSA Pop Report integration of Section 2.2 makes that conjecture directly testable. Before estimating the extended hedonic specification, we regress the signed mispricing residual from the original H6 (positive = market trades below fair value) on three candidate pop features with set fixed effects to absorb set-level calibration noise. Two features carry significant negative coefficients on the mispricing residual: `log(psa10_count)` and `log(total_graded)`. Gem rate carries a significant positive coefficient.

Two findings deserve emphasis.

First, the directions are opposite, and both are robust to the inclusion of set fixed effects.

Second, the negative sign on pop count contradicts the textbook supply-and-demand prediction. The textbook account would have high pop count depress the PSA-10 market price and therefore generate positive mispricing residuals (i.e., the model would over-predict). Instead we observe the reverse: high-pop cards trade systematically *above* the hedonic prediction. The interpretation we propose is that pop count is endogenous with respect to revealed demand. Cards that are valuable and well-known attract more grading submissions, so population growth and price level move together. Population is a liquidity / popularity proxy in the cross-section, not a supply proxy. Gem rate is normalized per card and is mechanically driven by holographic-finish robustness and print-run condition; it preserves the supply interpretation. We return to this distinction in §8.

3.2 Specification

The H6 v2 production specification adds two pop-derived regressors to the Model B specification of 2026-01:

$$\log(P_i) = \alpha + \beta_1 \log\_raw\_c_i + \beta_2 is\_sir_i + \beta_3 (log\_raw\_c \times is\_sir)_i + \beta_4 char\_top_i + \beta_5 char\_mid_i + \gamma_1 artist\_top_i + \gamma_2 smart\_divergence_i + \gamma_3 velocity\_z_i + \delta_1 \log\_psa\_total\_graded_i + \delta_2 psa\_gem\_rate_i + \Sigma_s \zeta_s Set_s + \Sigma_t \theta_t Type_t + \varepsilon_i$$

Estimation remains OLS with HC3 robust standard errors. To avoid extrapolation into thinly graded territory, we apply a training-time filter MIN_TOTAL_GRADED ≥ 30. Cards below the threshold are dropped from the fit. At inference time the same filter is applied: cards with `total_graded < 30` impute both pop regressors to the training-sample mean of those features. This preserves cross-sectional comparability without amplifying noise from cards with near-singleton grading samples.

3.3 Results

We re-estimate Model B and H6 v2 on a temporally aligned snapshot of the hedonic core. Table 2 reports the headline fit statistics.

Table 2. Comparison of Model B and H6 v2 on a temporally aligned snapshot. CV reported as the median across five random folds; LOSO is the average across all fourteen set holdouts.

Statistic	Model B (2026-01)	H6 v2 (this paper)	Δ
n (after filter)	358	343	-15
In-sample R^2	0.879	**0.914**	+0.035
In-sample Adj. R^2	0.870	0.907	+0.037
5-fold CV R^2	0.839	**0.873**	+0.034
5-fold CV median |%err|	17.5%	**15.8%**	-1.7 pp
LOSO CV R^2	0.786	**0.867**	+0.081
LOSO CV median |%err|	25.5%	**19.6%**	-5.9 pp

The new pop regressors enter with the expected signs. `log_psa_total_graded` carries a coefficient of approximately +0.137 ($p < 0.001$), indicating that a tenfold increase in graded population is associated with roughly a +37% premium over fair value, all else equal. `psa_gem_rate` carries a coefficient of approximately −1.589 ($p < 0.001$), indicating that a 10 percentage-point increase in gem rate (e.g., 30% → 40%) is associated with approximately a −15% reduction in fair value.

The improvement on leave-one-set-out cross-validation is particularly noteworthy. The 5.9 percentage-point reduction in LOSO median absolute percent error suggests that the pop variables capture cross-set variation in liquidity and condition that was previously absorbed only imperfectly by set fixed effects.

3.4 Economic interpretation

The opposing-signs structure carries direct interpretive content for the platform's user-facing labels.

A "Discount" label (positive mispricing) on a card with extreme pop count should be read with caution under the original H6; under H6 v2 the pop-count channel is absorbed into the fair value itself, and any residual discount or premium is conditioned on the population. A "Discount" label after H6 v2 deployment is, by construction, a deviation from fair value *that already accounts for graded supply and liquidity*.

We retain the H6 Risk Score (Section 7 of Baro 2026) as a complementary surface. Pop count alone has limited interpretive power as a risk signal because of the demand endogeneity documented in §3.1; eBay sales velocity and the smart-market divergence proxy continue to measure market thinness and hype-driven pricing instability that the level signal cannot.

3.5 Caveat on the residual-regression magnitudes

The residual-regression of §3.1 yielded coefficient magnitudes (on the order of -6.4 on `log(psa10_count)` and +89.6 on `gem_rate` in mispricing-percentage units) that are sensitive to a small number of high-leverage observations whose underlying PSA specification identifiers were, at the time of the original residual regression, cached pointing at non-English-language alternate specifications of the same cards. Re-resolution against the corrected specification catalog and re-running the residual regression on the cleaned sample reduces the `log(psa10_count)` coefficient magnitude substantially and renders it statistically insignificant at conventional levels, while leaving the `gem_rate` coefficient broadly intact.

The defensible source for the magnitude of the pop effect is therefore the production model's own coefficients (`log_psa_total_graded` = +0.137, $p < 0.001$; `psa_gem_rate` = −1.589, $p < 0.001$), not the residual-regression decomposition. The directional split — population as a liquidity proxy, gem rate as a supply proxy — is unchanged after correction.

---

4. H6 v2 PSA-9: A Parallel Specification at a Different Grade Tier

4.1 Motivation

The 2026-01 paper focused exclusively on PSA-10 prices. The PSA-9 market is materially deeper than the PSA-10 market in terms of population (typically 3–10× the volume of PSA-10 copies) and represents the marginal grade tier for many collectors who hold or trade outside the gem-mint cohort. A fair-value estimate at PSA-9 is therefore directly useful for collectors making cross-tier purchase or sale decisions.

4.2 Specification and results

We re-estimate the H6 v2 specification with `log(psa9_avg_price)` as the dependent variable, using PSA-9-specific sales metrics in place of their PSA-10 counterparts. Specifically, `smart_divergence` is recomputed as the absolute relative deviation between the PSA-9 smart-market price and the PSA-9 simple-sales-average price, and `velocity_z` is computed from PSA-9 daily sales volumes. The remaining regressors (raw price, rarity, character tiers, artist indicator, set FE, type FE, and the two pop regressors) are retained without modification.

A liquidity filter `psa9_count ≥ 10` is applied in addition to the inherited MIN_TOTAL_GRADED ≥ 30 filter. After both filters the panel contains 347 cards.

Table 3. H6 v2 specification at PSA-10 vs. PSA-9.

Statistic	PSA-10 (Section 3)	PSA-9 (this section)
n (after filters)	343	347
In-sample R^2	0.914	**0.920**
In-sample Adj. R^2	0.907	0.913
5-fold CV R^2	0.873	**0.905**
5-fold CV median |%err|	15.8%	**14.5%**
LOSO CV R^2	0.867	0.745
LOSO CV median |%err|	19.6%	24.3%

Two patterns are worth noting. First, the PSA-9 specification fits slightly tighter in-sample and on random k-fold CV, which is consistent with the PSA-9 market being more competitive and less driven by gem-mint scarcity tail effects. Second, the LOSO performance is materially weaker for PSA-9 (R^2 0.745 vs. 0.867 at PSA-10), suggesting that cross-set generalization at PSA-9 is more sensitive to set-FE fragility. We interpret this as evidence that PSA-9 prices are more strongly driven by within-set normative pricing dynamics — i.e., what comparable cards in the same set sell for at PSA-9 — than PSA-10 prices, which are anchored more directly by per-card scarcity premia.

4.3 Top-tier artist heterogeneity across grade tiers

The empirically identified top-tier artist set differs across grades. The procedure described in Baro (2026) §3.2 — identify artists with at least three cards in the sample whose median residual exceeds a threshold in log-price space — is applied separately to the PSA-9 and PSA-10 residuals.

The intersection of the two lists is approximately one-half of each. Two artists appear on both lists; two are PSA-10-specific and two are PSA-9-specific. The pattern is consistent with collector behavior diverging across grade tiers: some artists' work commands a particularly large premium at gem-mint condition (perhaps because the artwork is judged especially well-preserved or distinctive in flawless condition), while others command a premium more broadly across the grade spectrum.

We hold the specific composition of both lists proprietary, in line with the trade-secret framing of Baro (2026) §3.2.

4.4 PSA-9 / PSA-10 ratio diagnostic

For each card with both PSA-9 and PSA-10 prices observed, we compute the ratio PSA-9 / PSA-10. The median ratio across the panel is 0.256 (interquartile range 0.200 to 0.306). Thirty-two of 347 cards fall outside the plausible range [0.15, 0.95], typically because of thin volume on one side of the ratio. These observations are flagged for manual review rather than excluded from the fit, since most reflect real but idiosyncratic price discovery (e.g., a single high-priced PSA-9 sale in a market with few comparable transactions).

---

5. Out-of-Sample Validation

The 2026-01 paper reported only in-sample diagnostics. We now report out-of-sample fit along three axes and a preregistered walk-forward forward test whose binding evaluation point lies in the future.

5.1 Five-fold random cross-validation

We partition the 358-card panel into five random folds, fit Model B (the 2026-01 specification, *not* H6 v2) on four folds, and predict held-out fold prices. To control for in-fold leakage from data-driven feature construction, we re-derive within each fold: (i) the log-raw centering constant, (ii) the set fixed-effect baseline, and (iii) the empirical top-artist list. The held-out fold's set fixed effects are imputed as the training-fold mean.

Table 4. Five-fold random CV results for Model B.

Fold	n_test	R^2	Median |%err|
1	72	0.842	16.8%
2	72	0.851	17.9%
3	72	0.831	18.4%
4	72	0.844	16.5%
5	72	0.827	17.8%
Mean		**0.839**	**17.5%**

The 5-fold R^2 of 0.839 represents only a small reduction from the in-sample R^2 of 0.879, indicating that the model is not materially overfit to the training data.

5.2 Leave-one-set-out cross-validation

We then conduct a more stringent leave-one-set-out (LOSO) cross-validation, in which one of the fourteen sets is held out entirely and the model is fit on the remaining thirteen.

Table 5. LOSO CV results for Model B, by set.

Held-out set	n_test	R^2	Notes
Paldean Fates	11	0.961	Strong generalization
Surging Sparks	32	0.921	Strong
Prismatic Evolutions	29	0.913	Strong
Journey Together	17	0.892	Strong
(8 mid-range folds)	—	0.57–0.83	Within expected range
Shrouded Fable	15	0.054	Atypical theme, small set
Scarlet & Violet	10	−0.182	Structural artifact (FE baseline)
Mean (excluding baseline)		**0.786**	**Median |%err| 21.5%**

The mean LOSO R^2 of 0.786, computed excluding the structurally weak Scarlet & Violet baseline fold, demonstrates that the model generalizes to held-out sets at a level of fit only modestly below the in-sample performance.

Two folds underperform the rest. The Scarlet & Violet fold (the SV1 base set) is the omitted FE baseline in the production specification; removing it leaves the model without a calibration anchor and the resulting R^2 is structurally negative. The Shrouded Fable fold (n=15) is a small thematic spinoff set with limited cross-set comparables; we interpret this as an honest reflection of model limits on atypical sets.

5.3 180-day convergence backtest (October 2025 — April 2026)

We additionally test whether the original H6 model's fair-value estimates from October 31, 2025 predict observed market prices six months later, on April 29, 2026. The test uses no live API calls: all data are read from the platform's historical snapshot tables. After applying liquidity filters (PSA-10 sales count ≥ 10 at both endpoints, complete prediction record), the test sample contains 160 cards.

Table 6. 180-day convergence backtest results.

Metric	In-sample (today)	Out-of-sample (180-day)
n	358	160
Median absolute error	18.2%	29.4%
Median signed error	+1.8%	+22.8%
Within ±15%	40% of cards	28% of cards
Within ±30%	70% of cards	51% of cards
Within ±50%	88% of cards	73% of cards

The +22.8 percentage-point signed error over six months reflects a broad market-wide appreciation of the SV-era TCG segment during the test window, not a systematic bias in the model. When the cross-sectional mean drift is removed, idiosyncratic per-card volatility dominates the residual error structure.

The economically important reading of this result is that the H6 framework provides a cross-sectional relative-valuation tool, not a price-level forecast. The framework is analogous in spirit to traditional equity multiples such as price-to-earnings or price-to-book: useful for ranking comparables and risk-managing cohorts, but not a basis for absolute price prediction without exogenous information about market regime.

5.4 Walk-forward forward test (preregistered, 2026-05-01)

To address the cross-sectional / time-series distinction directly, we have preregistered a walk-forward forward test of the H6 v2 specification at multiple horizons. The preregistration was filed on 2026-05-01 and is included as supplementary material to this paper. Headline elements:

- **Test start:** 2026-04-27 (first daily snapshot under the production cron).
- **Anchor structure:** rolling anchor, single-asset trials indexed by `(card_id, anchor_date, horizon)`. Anchors are drawn at every daily snapshot and held to the corresponding horizon date.
- **Horizons:** canonical 7-day, 30-day, 90-day, and 180-day; diagnostic 1-to-4-day series behind a separate flag.
- **Cohort assignment:** per-anchor quintile cohorts assigned strictly within the anchor's snapshot. Q5 contains the largest discounts (market < fair value), Q1 the largest premia.
- **Test statistics:** Mann-Whitney U comparing Q5 vs. Q1 forward returns; Spearman rank correlation between mispricing rank and forward return; block-bootstrap confidence interval on Q5 − Q1 spread; per-cohort 5%-Value-at-Risk.
- **Decision rule (H1-binding):** at horizons 30, 90, and 180 days, all of (i) Mann-Whitney $p < 0.0167$ (Bonferroni-corrected at $\alpha=0.05$ across three horizons), (ii) bootstrap CI on Q5 − Q1 spread excludes zero, (iii) Spearman $\rho > 0$, and (iv) premium-cohort 5%-VaR < discount-cohort 5%-VaR.

The first H1-binding evaluation falls at 2026-07-26 (test start + 90 days). The 30-day evaluation falls at 2026-05-27. Results will be reported in a subsequent working paper.

5.5 Multi-anchor robustness sprint (2026-05-11)

A separate post-hoc robustness check, run after the §7 single-anchor analysis was published, addresses a direct concern with the §7 results: that the +60% Q5 vs. +43% benchmark headline depends on a single anchor date (2025-05-08) chosen ex-post. To stress that result, we re-fit the production raw-price XGBoost model at four independent anchors one month apart and measured 365-day forward returns from each:

Table 7. Multi-anchor Q5-Q1 spread sensitivity (mean returns, eligible cohort).

Anchor (T_0)	n eligible	Q1 mean	Q5 mean	Q5 − Q1 spread
2025-02-07	654	+7.6%	+52.3%	**+44.7pp**
2025-03-07	710	+21.2%	+38.8%	+17.6pp
2025-04-07	665	+44.9%	+35.1%	**−9.7pp**
2025-05-07	648	+63.7%	+40.5%	**−23.3pp**
Median across anchors				**+3.94pp**

The Q5 − Q1 mean return spread inverts systematically across the anchor sweep. At the earliest anchor it is +44.7 percentage points; at the latest it is −23.3 percentage points; the directional flip occurs between the 2025-03 and 2025-04 anchors and is monotonic. Median spread across the four anchors is +3.94 percentage points, effectively zero relative to the +44pp magnitude at the most favourable anchor. Three plausible mechanisms (none disambiguable from the cross-sectional data alone): a bull-market regime favouring established quality picks in Q1; forward-looking pop-count imputation bias in earlier anchors; and thin volume-feature history at early anchors. The §7 single-anchor result of +59% Q5 mean is the worst-of-anchors data point — at the 2025-05-08 anchor the spread is in fact negative — making the headline statistically fragile.

The same anchor sweep yields a sharply different result for risk-adjusted performance.

Table 7b. Multi-anchor risk-adjusted comparison (annualized).

Anchor	Q1 Sharpe	Q5 Sharpe	Q1 ann. vol	Q5 ann. vol	Q5 vol reduction
2025-02-07	0.24	**3.43**	56%	30%	−46%
2025-03-07	0.90	**2.23**	48%	30%	−37%
2025-04-07	2.37	**2.57**	29%	21%	−27%
2025-05-07	2.88	**3.29**	27%	17%	−39%
Median	**1.60**	**2.88**	**40%**	**24%**	**−39%**

In every anchor tested, the Q5 cohort exhibits a higher Sharpe ratio (2-10× higher than Q1), substantially lower annualized volatility (27-46% reduction), lower max drawdown, and higher Calmar ratio. This pattern survives even at the late anchors where the total-return spread inverts.

The sprint additionally segments the pooled cross-section by era, rarity, price bracket, and pop-count bracket. The Q5 − Q1 mean spread is strongly positive in Sun & Moon era (+23.5pp, $p<0.001$), pop_bracket <100 (+21.0pp, $p<0.001$), price_bracket $10–50 (+13.7pp, $p<0.001$), and several older-style rarities (Rare Holo GX +39.4pp, Rare Ultra +25.8pp, Rare Rainbow +9.6pp). It is strongly negative in Scarlet & Violet era (−33.6pp), Sword & Shield era (−25.1pp), Special Illustration Rare (−33.2pp), Illustration Rare (−14.7pp), and the $50–200 price bracket (−18.5pp). The model's signal lives in a sub-segment that is the *complement* of the segment on which the §7 single-anchor analysis is centred.

This sprint substantially weakens the total-return claim of §7 and substantially strengthens the risk-management claim. Sections §7 and §8 should be read with this in mind: the cross-sectional return spread is anchor-sensitive and not multi-anchor robust, while the cross-sectional volatility spread is anchor-invariant and robust at the 100% of-anchors level. The full sprint protocol, per-anchor outputs, stage-gate decisions, and caveats are documented in `findings/2026-05-11_alpha_validation_sprint.md`.

---

6. Raw-Price Extension: A Gradient-Boosted Tree Ensemble

6.1 Motivation and design constraints

Section 9 of Baro (2026) noted that the original hedonic specification was restricted to PSA-10 graded prices and did not directly address raw-market price formation. The Hedonix platform requires a raw-price fair-value estimate to surface alongside graded estimates for cards held in raw condition. We document here a separate model dedicated to this task.

The design constraints differ materially from the PSA-10 specification of §3. The raw market is heavily noisier than the graded market: TCGplayer raw quotes are vulnerable to bulk listings, undercut bots, and floor-of-market noise, particularly for cards quoted below $10. The dependent variable itself is therefore noisier than the PSA-10 average. In addition, raw cards lack a clean revealed-preference demand instrument analogous to PSA gem rate or grading volume — there is no per-card "raw grading population" because raw cards are by definition ungraded.

6.2 Methodology trail

We pursued a sequence of architectures, recorded here in part because two widely discussed feature classes returned null results that may be of independent methodological interest.

Attempt 1 — Reproduction of a published architecture. A publicly described raw-price model uses two regressors: a per-rarity "Pull Cost" and a three-component "Desirability Index" (character premium, artwork/hype rating, universal appeal). Reproducing the structural component on the original 358-card SV panel and on the expanded 2,622-card universe yields LOSO R^2 of 0.31 (SV-only) and 0.43 (expanded universe), respectively. The Pull Cost regressor is not statistically significant on our panel.

Attempt 2 — Large language model artwork scoring. We prompted Claude Sonnet 4.6 to rate 100 cards on five aesthetic dimensions (composition, pose, color, background, narrative). The resulting scores correlate with log raw price at coefficients below 0.30 in absolute value, indicating the scoring is free of price-recognition leakage. Adding the five scores to the model yields an in-sample R^2 lift of +0.003 and a five-fold CV lift of −0.15. The aesthetic dimensions captured by general-purpose vision-language models are not informative for raw-card pricing once rarity, character, and set are controlled for.

Attempt 3 — Google Trends as universal appeal proxy. We pulled 12-month Google Trends interest scores for 147 characters out of 300 attempted (the remainder were rate-limited). Adding the character-level interest score as a regressor yields an in-sample R^2 lift of +0.001 and zero LOSO lift.

Attempt 4 — Continuous character premium. Replacing the binary character-tier flags with a continuous score derived from per-set average price-rank of each character introduces estimation-variance problems for rare characters. With cutoff or Bayesian shrinkage applied to characters with insufficient observations, the continuous formulation matches the binary baseline cleanly but does not exceed it. We conclude that the binary character-tier representation already captures the predictive content of character identity.

Attempt 5 — Graded features as instruments. The original embargo against using PSA-derived features in a raw-price model proved methodologically unfounded. Raw and graded prices are simultaneously observable in cross-section; using graded population volumes and contemporaneous PSA prices as features for raw-price prediction is not temporal forward-look leakage. Adding graded volumes and gem rate yields a +0.10 LOSO improvement; adding contemporaneous PSA prices yields an additional +0.32. We retain the graded *volumes* in the production raw model but exclude graded *prices*, since the latter creates a mechanically derived "fair value" that moves with daily PSA price swings rather than providing the stable per-card anchor collectors expect.

Attempt 6 — eBay daily volume time-series. PPT's `priceHistory.volume` field exposes daily eBay sales counts per card over a 730-day window. Aggregating to lifetime 30-day, 90-day, and 730-day volume features yields a +0.27 LOSO lift over the bigger-panel baseline. This is the largest single architectural gain in the sequence. Substantively: revealed-preference signals (what people are actually buying) dominate stated-preference signals (LLM ratings, Google Trends).

Attempt 7 — Gradient-boosted trees. Replacing OLS with XGBoost on the same feature set yields a +0.07 LOSO lift. The improvement reflects non-linear interactions (e.g., "Charizard × Special Illustration Rare × freshly released") that linear specifications under-capture.

6.3 Final specification and results

The production raw-price model uses the following feature classes:

- Card metadata (rarity, character tier, era flags, artist).
- PSA population *volumes* (log total graded, log PSA-10 count, log PSA-9 count) — without PSA prices.
- eBay sales velocity (30-day, 90-day, 730-day rolling sums and standard deviations).
- Sealed-product premium per set (Elite Trainer Box price relative to release-time MSRP).
- Character-tier and franchise-tier classifications (Eeveelution, Starter, Legendary, Mythical, Mascot).

- Set-composition aggregates (rare-tier counts, set age in months).

Table 7. Final XGBoost raw-price model performance, honest leave-one-set-out cross-validation.

Sample	n	LOSO R^2	Median |%err|
Full universe	2,622	0.83	34%
Restricted to raw ≥ $10	1,141	0.74	30%
H6 hedonic core (SV SIR/IR only)	183	—	—

In-sample R^2 on the full universe is 0.97; the gap to LOSO 0.83 quantifies the cost of cross-set generalization in a non-parametric model.

6.4 Eligibility filter

Because the raw model is fit on a heterogeneous universe and uses a tree ensemble that can produce out-of-distribution predictions, we apply a hard eligibility filter at inference time. A card receives a raw fair-value reading only if all of the following hold:

- Raw market price ≥ $10.
- PSA-10 population > 0 (i.e., at least one PSA-10 in circulation).
- Total graded population ≥ 5.
- Predicted raw fair value < contemporaneous PSA-9 market price.
- Predicted raw fair value < 0.9 × contemporaneous PSA-10 market price.

The last two filters are sanity bounds — a raw card cannot rationally be worth more than its PSA-9 equivalent — and protect against tree-ensemble extrapolation. Cards failing the filter surface only the market price with no fair-value reading. Of the 2,622 cards in the universe, approximately 648 pass the filter at any given snapshot; the eligible-cohort fraction varies with market regime.

6.5 Plafond on raw-price modeling

The structural finding from the methodology trail of §6.2 is that LOSO R^2 of approximately 0.83 represents a near-plafond for raw-price modeling with the data classes available to us. The H6 v2 PSA-10 model exceeds this (LOSO 0.87) by exploiting PSA gem rate as an exogenous, auditable measure of supply scarcity — a feature with no clean analogue in the raw market. We expect further progress on raw-price fair-value modeling to require either (i) a revealed-preference demand instrument we do not currently collect (e.g., eBay sold-listings micro-data at the bid/ask level via the Marketplace Insights API), or (ii) a narrower cohort scope with reduced heterogeneity.

7. Live Performance: The Hedonix Index

7.1 Methodology

We track the live one-year total return of an equal-weighted top-quintile portfolio constructed from the raw-XGB-eligible cohort of §6.4. At each weekly rebalance date, the eligible universe is partitioned into quintiles by raw fair-value spread, and the top-quintile (Q5, most discounted) and bottom-quintile (Q1, most premium-priced) cohorts are recorded. Daily NAV is computed from per-card cumulative total return between rebalance dates, chained at rebalance points.

The genesis cohort uses 648 raw-XGB-eligible cards as of T_0 = 2025-05-08, partitioned by the contemporaneous fair-value spread. The benchmark for comparison is the full research universe (n=2,388 with valid T_0 basis), equal-weighted, with no eligibility filter.

A methodological note on portfolio aggregation: both Q5 and the benchmark are reported as the arithmetic mean of per-card returns. This corresponds to the return on an equal-weight portfolio rebalanced at each card's purchase. Alternative aggregations (median per-card return, long-short Q5 − Q1) are also computed but are reported separately to avoid mixing conventions.

7.2 365-day results

Over the period T_0 = 2025-05-08 to T_1 = 2026-05-08:

Table 8. Hedonix Index one-year live performance.

Cohort	n	Mean return	Median return
Q5 (most discounted)	130	**+59.0%**	+53.7%
Q4	129	+39.9%	+32.9%
Q3	130	+47.5%	+32.8%
Q2	129	+35.7%	+29.1%
Q1 (most premium)	130	+27.8%	+21.6%
Q5 − Q1 spread		**+31.2 pp**	+32.1 pp
Broad-market benchmark	2,388	+43.8%	—
Q5 − benchmark spread		**+15.2 pp**	—

The figures reported on the Hedonix marketing surface (Q5 +60%, benchmark +43%, spread +17 pp) round to one decimal place; Table 8 reports to the analytic precision of the underlying snapshot data.

7.3 T_0-trained model: controlling for anachronism

A natural concern with the §7.2 results is that the model used to assign cohorts was estimated on data partially contemporaneous with the test window. To control for this, we re-fit the XGBoost raw model on data anchored at T_0 = 2025-05-08 (i.e., using only

information available at T_0) and re-rank the cohorts. PSA pop counts, sales velocity, and sealed-product premia are imputed at present values (no historical PSA snapshot exists for 2025-05-08); volume features and set-age variables are correctly anchored at T_0.

Table 9. T_0-trained model: anachronism-controlled cohort returns.

Cohort	n	Median return	Mean return
Q5 (most discounted)	130	**+53.7%**	+59.0%
Q4	129	+32.9%	+39.9%
Q3	130	+32.8%	+47.5%
Q2	129	+29.1%	+35.7%
Q1 (most premium)	130	+21.6%	+27.8%

Table 10. Original vs. T_0-trained: which signals survive.

Metric	Production-trained (2026)	T_0-trained (2025)	Verdict
Card-level directional hit rate	66.7%	49.1%	Card-level direction is at chance once anachronism is removed
Spearman ρ (eligible cohort)	+0.326	+0.172	Halved
Spearman ρ (hedonic core)	+0.432	+0.050	Collapsed
Q5 − Q1 median spread	+49.4 pp	+32.1 pp	Survives, with 17 pp reduction
Mann-Whitney p (Q5 > Q1)	<0.0001	3.2×10^{-6}	Both significant

The reading is consistent across the table. Cross-sectional extreme-quintile separation survives the anachronism control with strong statistical significance. Card-level directional prediction does not survive: with anachronism removed, the model's directional accuracy on individual cards is indistinguishable from chance. The framework is informative at the portfolio / cohort level, not at the individual-card level.

We interpret this finding as direct empirical support for the platform's positioning as a risk-management tool rather than an alpha-signal generator. The cross-sectional fair-value framework is genuinely informative about which cohorts of cards trade at extreme discounts or premia relative to comparables, and these cohorts diverge in subsequent returns. The framework does not, however, predict which individual card will appreciate.

7.4 Residual concern: pop-count temporal alignment

The T_0-trained model of §7.3 still uses present-day PSA population counts as features, because the platform's PSA scraping infrastructure began accumulating snapshots only after T_0 and no historical population time-series exists for the test window. Population counts grew monotonically over the test year, so cards heavily-graded recently appear artificially "high-population" at T_0, which may bias the model's T_0 fair-value estimates. The direction of the resulting bias on the Q5 − Q1 spread is ambiguous: heavily-graded cards are typically popular cards that often kept rising, so the imputation could be either inflating or deflating the spread.

A clean closure of this concern requires a daily PSA scrape going forward, accumulating into a real historical population time-series. The walk-forward forward test of §5.4, anchored at T_0 = 2026-04-27 and running into 2027, will provide such data. We will re-run the convergence test on a point-in-time-clean panel at T_0 + 1 year (2027-05-08).

---

8. Discussion

8.1 Risk management vs. alpha generation

Together, the convergence backtest (§5.3), the LOSO cross-validation (§5.2), and the T_0-trained live results (§7.3) converge on a consistent reading: the H6 framework provides defensible cross-sectional fair-value rankings whose extreme cohorts diverge in subsequent returns. It does not provide point-in-time directional forecasts at the individual-card level. This is consistent with the platform's user-facing language — "Discount" and "Premium" labels indexing position relative to model fair value — and inconsistent with any framing that would conflate cross-sectional fair-value rank with directional forecast.

The H6 Risk Score, introduced as an attention layer in Baro (2026) §7 and computed from eBay sales velocity and the smart-divergence proxy, retains its role under H6 v2. Pop count, now absorbed directly into the fair value, is not a useful risk signal because of its endogeneity with respect to demand documented in §3.1. Velocity and smart-divergence continue to surface market thinness and short-term hype that the level signal cannot capture.

8.2 The pop-count endogeneity problem

The cross-sectional regression of §3.1 cannot disentangle the supply and demand effects of pop count without an instrument. Supply theory predicts higher pop → lower price (more copies in circulation), demand theory predicts higher pop → higher price (more grading submissions reflect more demand), and the cross-sectional sign reflects whichever channel dominates the equilibrium.

A clean separation requires either (i) a time-series identification strategy in which lagged changes in pop are used to predict subsequent changes in mispricing — this becomes

feasible once the weekly pop scraper accumulates 8–12 weeks of snapshots; or (ii) a quasi-experimental identification using exogenous shocks to grading supply (e.g., a documented PSA grading-window shift or a major reprint event). We treat both as research extensions for the H6 v3 / next-revision program.

8.3 Stated vs. revealed preferences

The methodology trail of §6.2 yielded a robust null on two stated-preference feature classes: LLM-based aesthetic ratings and Google Trends interest. Both are *a priori* plausible candidates for an aesthetic-quality or popularity proxy that would close the residual structure noted in Baro (2026) §6.2. Neither moves the model.

The successful alternative was a revealed-preference time-series feature: eBay daily sales volume. We interpret this as evidence that, in a market with active price discovery, revealed-preference signals dominate stated-preference labels. By the time an aesthetic judgment or trend score has been formed, the relevant demand has already been transacted at the going price; the score does not contain marginal information.

This pattern carries an operational implication for future research: investment in additional data engineering (revealed-preference instruments) is likely to dominate investment in additional data labeling (manual ratings, LLM scoring, survey).

8.4 Reprint risk

We do not resolve the reprint-risk omission identified in Baro (2026) §6.3. The serious analysis of reprint risk requires a multi-year time-series spanning one or more reprint events; the platform's accumulated history is currently insufficient. The bulk historical pull of §2.3 provides a foundation for retrospective reprint-event analysis on the SWSH and SM eras, but the relevant variation is concentrated in a small number of identifiable events (e.g., Charizard reprints in Crown Zenith; the Brilliant Stars reprinting of Surging Sparks-era promos) and is not yet quantified within our framework. Development of a reprint-risk module remains on the research roadmap.

8.5 Sealed-product premium

We have integrated a Set context surface on the platform that displays sealed-product premium (Elite Trainer Box current price relative to MSRP) per set as a user-facing signal. Sealed premium is not currently part of the H6 production model. A direct test on the original 14-set panel — replacing the set fixed effect with a `sealed_premium × reprint_status` interaction — yields a small in-sample improvement but a LOSO regression that is sensitive to a single high-leverage outlier (the 151 set, with sealed premium approximately 12.5× MSRP at the time of the test). With n = 14 sets the interaction coefficient is too unstable to commit to production. We expect to revisit this question once the panel expands beyond approximately 30 sets, which the universe expansion of §2 makes feasible in principle but which requires additional sealed-pricing infrastructure outside the scope of this paper.

8.6 The risk-reduction effect (post-§5.5 amendment)

The multi-anchor sprint of §5.5 reveals a structural finding that the original single-anchor analysis under-weighted: in all four anchors tested, the Q5 cohort delivers a Sharpe ratio approximately 2× that of Q1 (median 2.88 vs. 1.60 against a 4% risk-free reference) and annualized volatility approximately 40% lower (median 24% vs. 40%). This pattern survives anchors at which the total-return spread is *negative* — i.e., even when Q5 underperforms Q1 in level terms, Q5 still has a tighter risk profile.

Three implications follow.

First, the platform's user-facing framing as a "risk-management tool, not an alpha-signal generator" — already its positioning in operator documentation prior to this sprint — is empirically warranted in a sense stronger than originally claimed. The cross-sectional fair-value spread does not robustly predict cohort-level *return* across anchors; it does robustly predict cohort-level *risk*. This is a defensible product story, analogous in spirit to the role of book-to-market or earnings-yield as risk-style factors in the equities literature (Fama and French, 1992), where the factor identifies a return *premium* in expectation but is more reliably a *style-tilt* signal in any given window.

Second, the directional inversion documented in §5.5 (positive Q5–Q1 spread at early anchors, negative at late anchors) is consistent with a Fama-French-style interpretation in which the cohort signal captures a low-volatility characteristic that earns its premium over long horizons and through full market cycles but not necessarily within any single 365-day bull-market window. The H1-binding live walk-forward test of §5.4 should accordingly be amended to incorporate risk-adjusted decision criteria (Sharpe spread, VaR spread) alongside the existing return-spread criterion. Without that amendment, a successful framework risks failing the test on a return metric the data do not robustly support.

Third, the §6.2 sub-segment analysis suggests the segment in which the cohort logic does generate a return premium is *not* the segment on which the current product is centred (Sun & Moon era, older-style rarities, sub-$50 price points), and the segment on which the product is centred is precisely where the inversion is sharpest (Scarlet & Violet era, Special Illustration Rares, Illustration Rares). A coverage expansion backward into the older-era segments is therefore both (i) a feasible near-term operational change given the data are already staged in `historical.card_universe`, and (ii) a methodologically motivated step to align the product's deployment with the segment in which the model demonstrably performs.

8.7 Stale-price-bias amendment to §5.5 Sharpe ratios (2026-05-18 revision)

The Sharpe-ratio claims in §5.5 ("Q5 cohort delivers median Sharpe approximately twice that of Q1, 2.88 vs. 1.60") and the volatility claim ("annualized volatility approximately 40% lower, 24% vs. 40%") were computed from daily portfolio NAV returns annualized by the conventional `sqrt(252)` factor. A post-publication diagnostic carried out on 2026-05-18, documented in full at `findings/2026-05-18_stale_price_bias.md`, identifies these numbers as substantially inflated by stale-price / smoothing bias.

The PokemonPriceTracker `smartMarketPrice` series that powers the daily cohort NAV is itself a rolling-window aggregate of the underlying transaction stream per card. A real price shock today propagates into the smoothed series across multiple subsequent days; daily returns therefore split shocks into several smaller pieces, daily standard deviation captures only a fraction of the true move, and the `sqrt(252)` annualisation compounds the under-estimate. Empirically, the variance-inflation factor between daily and monthly aggregation on the live Hedonix Index NAV is approximately 1.9× on Q5, 1.91× on Q1, and 1.94× on the broad-universe benchmark — i.e., annualised volatility computed from monthly returns is roughly double the daily-returns-based annualisation. Sharpe ratios deflate by approximately the same factor.

The bias does *not* operate through serial autocorrelation of daily portfolio returns. Empirical AR(1) coefficients on portfolio-level daily returns are near zero on every cohort tested (Q1 −0.095, Q5 −0.125, benchmark −0.097 over the 1-year window); equal-weight aggregation across asynchronously-updated cards cancels the per-card smoothing artifact at the portfolio level. The variance under-estimate is structural to the daily-aggregation choice itself, not to an AR(1)-style autocorrelation.

Applying the empirically observed monthly-vs-daily inflation factor to §5.5 numbers:

	§5.5 reported	Corrected (monthly-frequency)
Q5 Sharpe (median across anchors)	2.88	**~1.55**
Q1 Sharpe (median across anchors)	1.60	**~0.86**
Q5 − Q1 Sharpe spread	+1.28	**+0.69**
Q5 annualised volatility	24%	**~46%**
Q1 annualised volatility	40%	**~76%**

The qualitative claim that the Q5 cohort delivers approximately twice the Sharpe of Q1 survives the correction (1.55 / 0.86 ≈ 1.80); the directional implication of §5.5 and the §8.6 product-positioning argument (the platform as a cross-sectional risk-management tool rather than a directional alpha-signal generator) is unchanged. The absolute Sharpe levels reported in the original §5.5 are not defensible without the monthly-frequency recomputation.

Three corollaries follow.

First, the production Sharpe widget on the application's portfolio surface, which until 2026-05-18 also used daily smartMarketPrice returns, has been migrated to monthly-period sampling at the same revision. User-facing Sharpe numbers will therefore drop by a comparable factor (approximately 1.5–2× depending on portfolio composition). This is a correction of methodology, not a degradation of the product.

Second, the §5.4 preregistered walk-forward test, with H1-binding evaluation in July 2026, was registered on daily-frequency cohort returns. The corrected Sharpe / VaR-spread decision criteria proposed in §8.6 must be implemented at monthly frequency to be free of

the bias documented here. The preregistration document will be amended accordingly before the H1-binding evaluation.

Third, the methodology section of a subsequent revision (Working Paper 2026-04 or later) should adopt the monthly-frequency Sharpe as the primary metric, with daily-frequency results reported only as a contrasting diagnostic. A complementary research path — daily snapshotting of the per-card median sale price into a separate time series, less smoothed than `smartMarketPrice` — has been started on 2026-05-18; once approximately 60–90 days of independent median-return history accumulate, a third-source cross-check of the corrected Sharpe will become possible.

For full diagnostic detail, including the multi-frequency vol-inflation table, the AR(1) null result, and the methodological reason why the `(max − min) / median` half-spread is *not* a tradable bid-ask cost despite the original 3a design intent, see `findings/2026-05-18_stale_price_bias.md`.

---

9. Conclusion

This paper extends the hedonic pricing framework of Baro (2026) along five dimensions. First, the integration of PSA Population Report data lifts in-sample R^2 from 0.879 to 0.914 and, more substantively, raises leave-one-set-out cross-validation R^2 from 0.79 to 0.87 — the cross-set generalization gain that matters for production deployment. The two new regressors carry opposite signs and tell economically distinct stories. Second, a parallel PSA-9 specification yields broadly comparable cross-sectional fit and surfaces cross-grade artist heterogeneity that may itself reward further study. Third, the universe expansion from 360 to 2,635 cards enables a non-parametric raw-price model whose performance is at the structural plafond reachable with publicly-derivable features; the methodology trail leaves a clean null on two widely discussed feature classes (LLM artwork ratings, Google Trends) and a strongly positive result on a third (revealed-preference sales velocity). Fourth, out-of-sample validation along four axes — random k-fold, leave-one-set-out, 180-day convergence, and a 365-day live track — converges on a single reading: the framework is informative cross-sectionally at the cohort level and substantially less informative directionally at the individual-card level.

The fifth dimension is the post-publication multi-anchor robustness sprint of §5.5, conducted after the §7 single-anchor live results were initially circulated. This sprint stress-tests the §7 +60% headline against four independent anchor dates and reveals that the cross-sectional *return* spread is not multi-anchor robust: median spread across four anchors is +3.94 percentage points, and the spread inverts directionally between the 2025-03-07 and 2025-04-07 anchors. The single-anchor 2025-05-08 result on which §7 is centred turns out to be the *worst* anchor in the four-point sweep — at that anchor, the spread is in fact negative. The cross-sectional *risk* spread, in contrast, is robust: in 100% of anchors tested, the Q5 cohort exhibits roughly twice the Sharpe ratio and ~40% lower annualized volatility than the Q1 cohort. The §5.5 amendment to §8 (now §8.6) reads this finding as direct empirical support for the platform's positioning as a risk-management tool: the

framework provides defensible cohort-level risk-style screening even at anchors where it does not provide defensible return-style screening. The §6.2 sub-segment analysis additionally identifies that the segment in which the model does generate a robust return premium (older-era cards, older-style rarities, sub-$50 price points) is the complement of the segment on which the current product is centred — motivating a coverage-strategy realignment toward the older eras already staged in the historical schema.

Five open items structure the H6 v3 / next-revision research agenda: the temporal identification of the pop-supply versus pop-demand channels; a quantitative reprint-risk module; the integration of sealed-product premium as a set-level demand index once the analytic panel expands; an amendment to the preregistered walk-forward test of §5.4 to incorporate risk-adjusted decision criteria (Sharpe spread, VaR spread) alongside the existing return-spread criterion; and a deliberate coverage expansion to align the product's deployment universe with the segments in which §5.5 demonstrates a robust return premium. Subsequent working papers will document each as the underlying data accumulate. The walk-forward forward test preregistered in May 2026 will provide the first point-in-time-clean readout of cross-sectional cohort separation at canonical horizons of 30, 90, and 180 days; its H1-binding evaluation falls in July 2026.

References

Baro, P. (2026). *A hedonic pricing model for graded Pokémon trading cards: Evidence from the Scarlet & Violet era (2023–2026)*. Hedonix Research Working Paper No. 2026-01.

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Dimson, E., and Spaenjers, C. (2014). The investment performance of emotional assets. In V. A. Ginsburgh and D. Throsby (Eds.), *Handbook of the Economics of Art and Culture, Vol. 2* (pp. 521–549). Elsevier. https://doi.org/10.1016/B978-0-444-53776-8.00009-9

Mandel, B. R. (2009). Art as an investment and conspicuous consumption good. *American Economic Review*, 99(4), 1653–1663. https://doi.org/10.1257/aer.99.4.1653

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 8748–8763.

Renneboog, L., and Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. *Management Science*, 59(1), 36–53. https://doi.org/10.1287/mnsc.1120.1580

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55. https://doi.org/10.1086/260169

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. https://doi.org/10.2307/1912934

About the Author

Philipp Baro is a Bachelor's student in Frankfurt, Germany, working on Hedonix Research. He built the H6 hedonic engine, the multi-tier model ecosystem (PSA-10 / PSA-9 / raw XGBoost), and the multi-anchor validation framework documented in this paper. Hedonix Research is independent, self-funded, and has no outside team. Correspondence: philipp@hedonix.tech.

About Hedonix Research

Hedonix Research is an independent quantitative research project applying econometric methods (hedonic pricing, cross-sectional risk analytics, gradient-boosted ensembles) to the secondary market for graded Pokémon trading cards. Working papers are released periodically to document the methodologies underlying the H6 Hedonic Engine and to engage with the broader academic literature on collectible-asset pricing.

The views expressed in this working paper are those of the author and do not constitute investment advice. Past performance of the H6 Hedonic Engine in cohort-level validation tests does not guarantee future results, and no part of this paper should be read as a recommendation to buy or sell any specific card. The Hedonix platform is in private beta at hedonix.tech.