Two Heads Are Better Than One: t-Statistics and Monotonicity in the Factor Zoo

Jiawei Fan*

Brandeis University

Abstract

This paper shows that two in-sample diagnostics used to judge factor quality—the conventional t-statistic and monotonicity—capture distinct information and enhance predictive power when combined. The two measures are nearly uncorrelated in-sample ($\rho = -0.08$). Factors with the highest in-sample summed rank across both dimensions ($\geq$70th percentile) deliver 33.4/34.2 bps/month in out-of-sample return/alpha, while factors with the lowest summed rank ($\leq$30th percentile) deliver only 10.2/10.9 bps/month. High-summed-rank factors also perform better in real-time trading: a strategy built on the top 30% by summed rank dominates on a certainty-equivalent basis at every risk-aversion level considered and continues to hold up once trading costs and market impact are priced in. Finally, the combined strategy results in higher diversification in factor rotation.

*Jiawei Fan is from the School of Business and Economics at Brandeis University.

Imagine a journal editor receives a manuscript claiming a new return-predictive factor—validated with a reported t-statistic using all available historical data. She must decide whether the finding is likely to survive out-of-sample—so must a portfolio manager, before allocating capital to it. Researchers have long recognized that many asset-pricing factors weaken once they venture beyond their discovery sample: McLean and Pontiff (2016) document that mean factor returns drop 26% in out-of-sample tests and 58% after journal publication.

Among many responses to this problem, I consider two distinct dimensions along which a factor's robustness can be tested: the t-statistic and monotonicity, and show that predictability improves once both are taken into account. On one hand, since Harvey et al. (2016), a multiple-testing framework—whose underlying central measure is the t-statistic—has been widely adopted and has become the leading approach to testing factors' replicability (Hou et al., 2020; Chen and Zimmermann, 2022; Jensen et al., 2023). On the other hand, the concept of monotonicity, introduced by Patton and Timmermann (2010), has served as a formal statistical test for monotonic implications, but has never, to the best of my knowledge, been applied as a factor robustness test at the scale of the t-statistic-based literature above.

The t-statistic is, at its core, a signal-to-noise ratio, and is informative from both frequentist and Bayesian perspectives. Yet it has one notable limitation: the standard top-minus-bottom t-statistic relies only on the two extreme portfolios, discarding information contained in the intermediate ranks and thus failing to fully exploit the cross-sectional structure of returns (Patton and Timmermann, 2010). Monotonicity, by contrast, captures the "shape" of the return structure across the full range of sorted portfolios—a pattern implied by many financial theories, not merely a statistical curiosity—making use of the entire cross-section rather than just its extremes. It therefore serves as a natural, essential complement to the t-statistic.

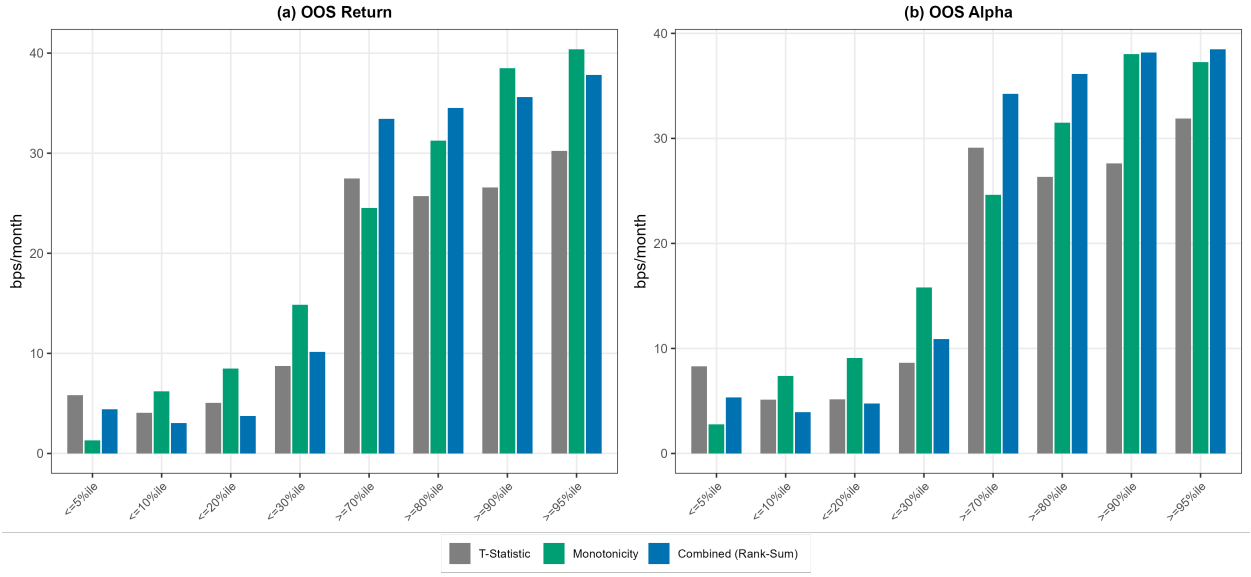
Rather than a formal statistical test, this paper treats monotonicity as a non-parametric, continuous measure bounded between 0 and 1, designed for ranking and easy implementation. I measure a factor's monotonicity as the share of months in which quintile portfolio returns rise (or stay flat) steadily from the lowest to the highest portfolio—an intuitive "fraction-monotonic" score. There may well be alternative, and possibly better, measures for capturing this kind of "shape" characteristic; I deliberately keep this one simple, intended to capture the idea rather than to be the last word on it. A useful direction for future research would be to develop a more robust measure that captures statistically and economically meaningful monotonic structure.

Figure 1 summarizes the paper's main cross-sectional result. Panel (a) plots each measure's mean out-of-sample return and panel (b) its mean out-of-sample CAPM alpha, for factors grouped into eight bins by their in-sample percentile rank on each of three mea-

sures: the t-statistic, monotonicity, and their rank-sum ("combined"). In the bottom four bins ($\leq$5th through $\leq$30th percentile), all three measures earn little (1–15 bps), confirming that weak in-sample signals carry correspondingly little out-of-sample value; monotonicity is, if anything, the *highest*-earning of the three in every low-signal bin except the very bottom one, so a low monotonicity score is not itself strongly punishing. The pattern reverses in the top four bins ($\geq$70th through $\geq$95th percentile): all three measures climb steadily, but the two constructed measures pull ahead of the t-statistic and trade the lead between them depending on the outcome. For out-of-sample *return*, monotonicity is the strongest performer at the two most extreme bins, peaking at 40.4 bps at the $\geq$95th percentile (versus combined's 37.8 bps and the t-statistic's 30.2 bps); for out-of-sample *alpha*, combined is the strongest performer at every top-tail bin, peaking at 38.5 bps at the $\geq$95th percentile (versus monotonicity's 37.3 bps and the t-statistic's 31.9 bps). Combined is, in this sense, the more consistent discriminator of the three: it leads or ties for the lead in six of the eight top-tail bin-outcome combinations, while monotonicity's edge is concentrated in raw return at the very extreme percentiles and the t-statistic never leads beyond the $\geq$70th bin. The in-sample cross-sectional correlation between the t-statistic and monotonicity is essentially zero ($\rho = -0.08$), consistent with the two measures capturing distinct dimensions of factor quality.

Figure I
Avg. OOS Return and Alpha by In-Sample Signal Percentile

This figure plots, for each of three in-sample measures — the t-statistic, monotonicity, and their rank-sum ("combined") — the mean out-of-sample return (panel a) and mean out-of-sample CAPM alpha (panel b) of factors falling in each of eight percentile bins of that measure: $\leq$5th, $\leq$10th, $\leq$20th, $\leq$30th, $\geq$70th, $\geq$80th, $\geq$90th, and $\geq$95th. The t-statistic is based on the top-minus-bottom factor return spread; monotonicity is the fraction of monotonic months over the total number of months (Section I.D); combined is the rank-sum of the two, computed once cross-sectionally over the full factor universe. The in-sample and out-of-sample periods are defined by each factor's original sample period (Section I.E). Sample: January 1926 through December 2024, $N = 150$ characteristics, VCAP (capped value-weighted quintile portfolios, NYSE non-microcap breakpoints; the only weighting scheme used throughout this paper, see Section I.C).



The predictive power of the t-statistic, monotonicity, and their rank-sum is further tested with cross-sectional regressions in Section II.A. Both the t-statistic and monotonicity predict out-of-sample returns and alphas on their own, and continue to do so essentially unchanged when entered together. Combining the two measures into a single rank-sum outperforms either alone, achieving the strongest fit of any specification and confirming that monotonicity contributes information the t-statistic cannot replace.

In the time series, the same complementarity carries over to monthly portfolio performance. I compare four strategies built from a common publication-based investable universe: trading all published factors, trading the top 30% ranked on up-to-date t-statistics, trading the top 30% ranked on up-to-date monotonicity, and trading the top 30% ranked on the sum of the two measures' ranks ("combined"). The low overlap between the t-statistic- and monotonicity-selected portfolios confirms that combining the two signals exploits genuinely complementary information rather than averaging away noise. While the t-statistic strategy has the highest Sharpe ratio (0.77, versus 0.71 for combined), the combined strategy dominates on a certainty-equivalent basis: across every risk-aversion level considered, it delivers

the highest certainty-equivalent return of the four strategies, including at $\gamma = 10$.

A long-only strategy can sometimes be more attractive than a long–short implementation, depending on how much of a strategy's performance is concentrated in a single leg. In long/short leg decompositions, this asymmetry is pervasive but not universal. All-factor's long and short legs are both individually significant and close in magnitude, but the t-statistic strategy's spread significance comes only from its two legs adding constructively, since neither leg is significant alone. Monotonicity's and the combined strategy's performance, by contrast, are concentrated almost entirely in the long leg: their short legs are statistically indistinguishable from zero and, in certainty-equivalent terms, value-destroying for risk-averse investors.

The highest Sharpe ratio, from the top-30% t-statistic strategy, does not survive once trading costs and frictions are priced in. Under a simple bid-ask-spread-and-borrow-fee cost schedule, its ranking inverts: the t-statistic strategy has the highest gross Sharpe (0.77) but the lowest net Sharpe of the four strategies (0.11–0.19) across every borrow-fee schedule considered, because its higher turnover consumes more of its edge. The combined strategy holds up best net of costs (0.23–0.29), and monotonicity's net Sharpe lands close to the t-statistic strategy's despite starting from a much lower gross Sharpe. A market-impact cost model shows the same inversion.

Traded factors are more diversified in the combined strategy than in either of its two parents. Examining which characteristics each strategy actually trades shows that the t-statistic and monotonicity strategies concentrate heavily in different, economically distinct clusters: the t-statistic strategy is dominated by INVESTMENT characteristics (asset-growth and overinvestment measures, consistent with a q-theory risk-based account), while monotonicity concentrates in TRADING FRICTIONS and VALUE characteristics (beta, idiosyncratic volatility, and spread measures more closely associated with limits-to-arbitrage proxies than with a rational risk story); the two strategies' top-20 rosters overlap in only 2 of 20 characteristics. The combined strategy's top 20 is not simply the union of its two parents — most of its members appear in neither parent's own top tier — and its family composition is markedly less concentrated: its largest single-family share is 35%, versus 70% and 45% for the t-statistic and monotonicity strategies, respectively.

The paper proceeds as follows. Section I describes the data and methodology. Section II presents empirical results. Section III examines which characteristics each strategy trades and what that implies about mechanism. Section IV concludes.

I. Data and Methodology

A. Data Sources

Return data come from the CRSP U.S. monthly stock file, and accounting data come from Compustat. The sample spans January 1926 through December 2024. Excess returns are measured relative to the U.S. Treasury bill rate. To reduce the influence of data errors, I winsorize monthly stock returns at the 0.1th and 99.9th percentiles. I incorporate delisting returns from CRSP whenever available; when a delisting return is missing and the delisting is performance-related, I assign a return of -30%, following Shumway (1997).

B. Characteristic Universe and Family Taxonomy

I construct 153 stock characteristics following Jensen et al. (2023), comprising 94 accounting-based and 59 price-based characteristics. For accounting-based variables, I assume accounting information becomes available four months after the fiscal period end; valuation ratios use the most recently available price.

Each characteristic is assigned to one of six economic families — INTANGIBLES, INVESTMENT, Momentum, Profitability, TRADING FRICTIONS, and VALUE — a literature/theory-based taxonomy (not derived from the characteristics' own return correlations) used throughout this paper: as a control in the cross-sectional tests (Section II.A), to describe which economic themes each strategy trades (Section III), and to summarize the characteristic universe itself (Table I, Panel A).

C. Portfolio Construction and VCAP Weighting

For each characteristic in each month, I sort eligible stocks into five quintile portfolios. The factor return is the top-quintile return minus the bottom-quintile return — a zero-net-investment long-short spread — where the factor is long (short) the quintile the original study identifies as having the highest (lowest) expected return. Quintile breakpoints are computed from NYSE-listed, non-microcap stocks only, then applied to sort the full eligible universe, rather than from the full eligible sample itself, which would let the large mass of illiquid microcap stocks distort the cutoffs. A factor-month requires at least five stocks in both the long and short legs; a characteristic requires at least 60 non-missing monthly observations to enter the sample. Turnover is the sum of the absolute changes in a stock's portfolio weight from one rebalance to the next, where the prior month's target weight is first drifted forward by its realized return before differencing ("drifted pre-trade weights"), rather than differenced as raw target weights.

I weight each quintile portfolio by market equity, winsorized at the NYSE 80th percentile — **capped value weight (VCAP) weighting**, the only weighting scheme reported in this paper. Each stock's market equity is capped at the contemporaneous NYSE 80th percentile breakpoint before being used as a portfolio weight (each leg's weights sum to one independently): stocks below the cap keep their actual market equity, so tiny stocks still receive correspondingly tiny weights, while any stock above the cap is weighted as if it were exactly at the 80th percentile, so no single mega-cap stock can dominate a portfolio no matter how large it actually is. This construction is designed to produce portfolios that are tradable yet balanced, avoiding the failure mode of full-universe value weighting, in which a handful of the very largest firms dominate the cross-section.

D. Monotonicity and Factor-Level Statistics

A month is defined as monotonic if quintile portfolio returns are non-decreasing from the bottom to the top quintile. For factor i with quintile portfolio returns $r_{i,t}^{(1)}, \ldots, r_{i,t}^{(5)}$ in month t, I define the monotonicity measure over a given period $\{1, \ldots, T_i\}$ as

$$\text{Monotonicity}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \mathbf{1}\left\{ r_{i,t}^{(1)} \leq r_{i,t}^{(2)} \leq r_{i,t}^{(3)} \leq r_{i,t}^{(4)} \leq r_{i,t}^{(5)} \right\}.$$

I compute this measure two ways. A *static* version is estimated once over a fixed window (e.g., a characteristic's in-sample period) and used for the cross-sectional tests in Section II.A and in Table I. An *up-to-date* (expanding-window) version is recomputed every month using only data available through that month; this recursive series, together with an analogous up-to-date t-statistic, is what Sections II.B–III use to decide which characteristics a strategy trades in real time.

I also report each factor's CAPM alpha: a static, factor-level OLS regression of the factor's own return series on the US value-weighted market excess return, estimated once over each of the sample windows defined in Section I.E (not a recursive monthly series). Section II additionally reports each portfolio strategy's alpha under three richer benchmark models — Fama-French three-factor, Fama-French five-factor plus momentum, and the Hou-Xue-Zhang q-factor model — detailed in Table II's and Figure 2's own notes.

E. Sample Periods and Strategy Construction

I define four sample windows for each characteristic. The *in-sample* period is the original study's own sample window. The *out-of-sample* period runs from the end of the in-sample period through the most recent month in my data. The *post-publication* period begins in

the characteristic's publication year. The *full-sample* period spans the characteristic's entire available history. Table I reports summary statistics under each window.

Sections II.B, II.C, II.D, and III restrict further to a *live-trading* sample: because characteristics are published at different times, I begin this sample in January 2000, by which point 35 characteristics are already published, ensuring a sufficiently large and stable investable universe from the outset; the live-trading sample runs through December 2024 (300 months).

Within the live-trading sample, I compare four portfolio strategies, each rebalanced monthly using only each characteristic's up-to-date t-statistic and monotonicity as of that month: (1) **All-factor**, trading every eligible characteristic; (2) **Top 30% t-stat**, trading the 30% of eligible characteristics with the highest up-to-date t-statistic; (3) **Top 30% monotonicity**, trading the 30% with the highest up-to-date monotonicity; and (4) **Combined**, trading the 30% with the highest summed rank across the two measures. A strategy's monthly return is the equal-weighted average, across its currently-selected characteristics, of each characteristic's own VCAP factor return.

Table I: Summary Statistics

Table I presents summary statistics for the data and variables used throughout the paper. I construct 153 stock characteristics, of which 149 are ever-eligible for the post-2000 live-trading sample used in Sections II–III; characteristics were published between 1973 and 2020 (median 2006), with 35 published before 2000 and 118 in 2000 or later. Panel A reports the family composition of the full 153-characteristic universe. Panel B reports the mean, across all factors, of each factor-level statistic, separately for the full sample and the in-sample, out-of-sample, and post-publication windows. Panel C reports the distribution of the key variables at the factor-month level over the full sample period.

Panel A: Family Breakdown (of 153)

Family	N
INVESTMENT	33
INTANGIBLES	31
Profitability	30
TRADING FRICTIONS	29
VALUE	18
Momentum	12

Panel B: Mean of 153 Factors, Across Sample Periods (VCAP)

	Full Sample	In-Sample	Out-of-Sample	Post-Publication
Return (bps)	30.29	36.17	22.29	19.37
CAPM Alpha (bps)	31.52	37.11	23.44	20.28
T-Statistic	2.60	2.53	1.04	0.83
Monotonicity	0.084	0.083	0.086	0.087

Panel C: Descriptive Summary at Factor-Month Level (VCAP, Full History)

Variable	N	Mean	SD
Return (bps)	140,635	28.37	378.39
Long Leg (bps)	140,635	16.72	611.21
Short Leg (bps)	140,635	11.65	589.19
Avg. Size, Traded Stocks ($000s)	144,985	2,198.2	3,941.9
Turnover	144,985	0.433	0.493
Up-to-Date T-Statistic	93,795	2.68	2.39
Up-to-Date Monotonicity	95,379	0.076	0.054

Variable	Min	5th	10th	25th	Median	75th	90th	95th	Max
Return (bps)	−7,089	−469	−307	−123	24.00	177	371	549	7,062
Long Leg (bps)	−9,101	−885	−655	−327	19.29	358	683	923	7,858
Short Leg (bps)	−9,206	−851	−626	−310	16.17	332	639	869	9,035
Avg. Size, Traded Stocks ($000s)	4.1	63.6	108.1	226.4	498.3	2,617.4	6,196.8	9,641.6	63,071.7
Turnover	0.00	0.05	0.06	0.09	0.173	0.613	1.338	1.610	2.00
Up-to-Date T-Statistic	−4.68	−0.65	−0.19	0.98	2.37	4.25	6.00	7.11	9.63
Up-to-Date Monotonicity	0.00	0.016	0.023	0.037	0.060	0.103	0.158	0.185	0.426

II. Empirical Results

A. Cross-Sectional Factor Performance

Table II reports the cross-sectional horse race between in-sample t-statistics, monotonicity, and their rank-sum in predicting out-of-sample returns and alphas, progressively across five specifications; see the table description for the full detail. All ten columns include the same five controls, so the progression isolates two questions: whether adding the second predictor changes either coefficient (REG1–3/REG6–8), and whether family fixed effects and the cluster-robust inference they require change the picture (REG3 to REG4/REG8 to REG9).

Univariately (REG1/REG6), the t-statistic is a strong predictor of both out-of-sample return ($t = 4.07$) and alpha ($t = 4.04$); monotonicity alone (REG2/REG7) is weaker but still significant (return $t = 2.71$, alpha $t = 2.31$). Adding both together (REG3/REG8) leaves each coefficient's magnitude and significance essentially intact — consistent with the near-zero correlation between the two measures documented throughout this project, each carries information the other does not.

Adding family fixed effects to the joint specification (REG4/REG9) is where the two measures diverge: monotonicity's coefficient survives, weakened but still significant (return $p = 0.033$, alpha $p = 0.052$), while the t-statistic's does not (return $t = 1.76$, $p = 0.139$; alpha $t = 1.74$, $p = 0.143$) — family composition evidently absorbs a meaningful share of the t-statistic's raw explanatory power but much less of monotonicity's. Replacing the two separate predictors with their summed rank, still under family fixed effects and cluster-robust inference (REG5/REG10), performs better than either predictor alone under this same demanding specification: the summed rank is significant at the 5% level for both outcomes ($t = 3.18$ for return, $t = 3.16$ for alpha) and achieves the highest R^2 of any of the ten columns (0.298 for return, 0.309 for alpha) despite collapsing two variables into one.

Table II: OOS Returns and Alphas on In-Sample T-Statistic and Monotonicity

Table II reports the cross-sectional horse race between in-sample t-statistics, monotonicity, and their rank-sum (the "combined" construction defined in Section I.E, computed once cross-sectionally over the full characteristic universe), all standardized, in predicting out-of-sample factor returns and out-of-sample CAPM alphas (Section I.D). All ten specifications include the same five controls (in-sample precision, size, and liquidity: n IS months, average stocks per portfolio, IS return volatility, IS turnover, IS market cap). REG1–REG5 regress out-of-sample return (`ret_oos_vcap`) on the t-statistic alone, monotonicity alone, both together, both together plus the six-category family fixed effects defined in Section I.B, and the summed rank alone plus family fixed effects, respectively. REG6–REG10 repeat the same five specifications for out-of-sample CAPM alpha (`alpha_oos_vcap`). t-values, shown in parentheses below each coefficient, use *mixed inference*: REG1–REG3 and REG6–REG8 (no family fixed effects) report plain OLS standard errors; REG4–REG5 and REG9–REG10 (with family fixed effects) report cluster-robust standard errors, clustered by family (6 clusters, $t(5)$ reference distribution per Cameron-Gelbach-Miller small-cluster guidance), since that is exactly where family-level dependence becomes mechanically relevant to inference. The Family FE and SE Type rows state which apply to each column. Constants and control/family-fixed-effect coefficients are omitted for brevity.

	OOS Return					OOS Alpha				
VARIABLES	REG1	REG2	REG3	REG4	REG5	REG6	REG7	REG8	REG9	REG10
In-Sample t-stat	9.05***		9.68***	7.29		9.55***		10.14***	7.77	
	(4.07)		(4.49)	(1.76)		(4.04)		(4.38)	(1.74)	
In-Sample monotonicity		7.04***	8.02***	8.39**			6.41**	7.44***	8.04*	
		(2.71)	(3.28)	(2.91)			(2.31)	(2.84)	(2.53)	
Summed Rank (Combined)					10.70**					11.03**
					(3.18)					(3.16)
Family FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
SE Type	OLS	OLS	OLS	Cluster	Cluster	OLS	OLS	OLS	Cluster	Cluster
Observations	150	150	150	150	150	150	150	150	150	150
R^2	0.146	0.093	0.206	0.290	0.298	0.156	0.094	0.201	0.298	0.309

*t-values in parentheses; see SE Type row for OLS vs. cluster-robust. *, **, *** denote significance at the 10%, 5%, 1% levels, based on the corresponding OLS or cluster-robust p-value. $N = 150$ characteristics (a small, consistent drop from the full 153-characteristic universe, not further investigated). Sample: January 1926 through December 2024, monthly U.S. stock return data from CRSP.*

B. Post-Publication Portfolio Performance

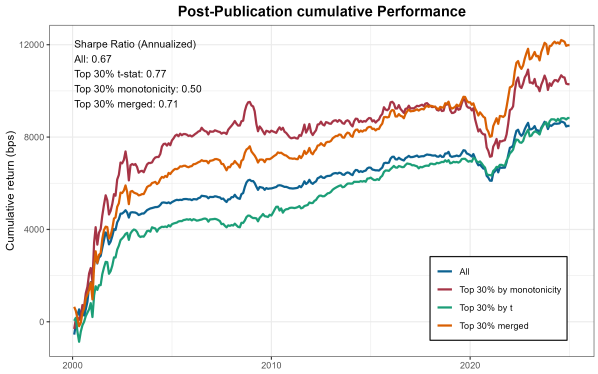
I examine factor portfolio performance after publication, over the live-trading sample and across the four strategies defined in Section I.E. As shown in Figure 2, all four strategies generate positive post-publication cumulative returns. Although the monotonicity strategy has a lower Sharpe ratio than the other three strategies (0.50, versus 0.67 for all-factor, 0.77 for t-statistic, and 0.71 for combined — Figure 2(c)), it nevertheless delivers persistently stronger cumulative returns over time than all-factor and t-statistic, ending only slightly below the combined strategy. The time-series correlation between t-statistics and monotonicity is persistently low, with a mean of -0.16, and the mean overlap share between the two selected portfolios is only 0.06. This indicates that the factors selected by the t-statistic strategy and the monotonicity strategy are largely distinct. As a result, combining the two signals allows

the portfolio to exploit complementary information. Consistent with this interpretation, the merged strategy delivers the highest cumulative return for most of the sample, though monotonicity alone closes much of the gap by the final years. Panel (c) of Figure 2 reports the full risk/performance battery for the four strategies: mean return and its Newey-West (HAC) t-statistic, alpha under four benchmark models (CAPM, Fama-French three-factor, Fama-French five-factor plus momentum, and the Hou-Xue-Zhang q-factor model) with their own HAC t-statistics, the annualized Sharpe ratio, and the certainty-equivalent annual return under mean-variance preferences at four risk-aversion levels. Every strategy's mean return and every alpha are HAC-significant at conventional levels, with one exception: monotonicity's alpha weakens to marginal significance ($p \approx 0.06$–0.07) once the FF5+momentum or q-factor controls are added, unlike the other three strategies, whose alphas stay significant under every model. On raw mean return and CE utility, the combined strategy dominates throughout (highest mean return, highest CE at every γ shown); on Sharpe, t-statistic dominates. Monotonicity is the highest-mean-return, highest-volatility strategy of the four.

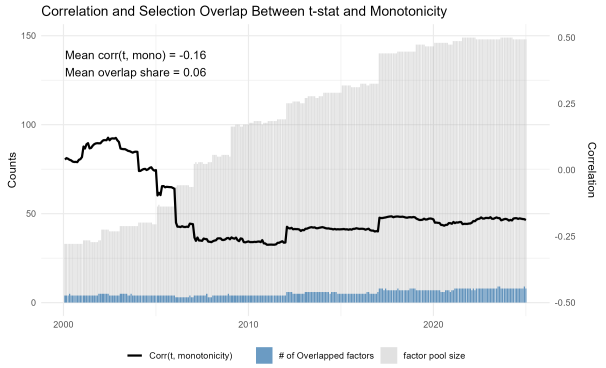
Figure II
Post-Publication Portfolio Performance across Strategies

Panel (a) plots cumulative returns from 2000 onward, in basis points, for four strategies: trading all published factors; trading the top 30% of factors ranked by up-to-date t-statistics; trading the top 30% ranked by up-to-date monotonicity; and trading the top 30% ranked by the sum of the t-statistic rank and monotonicity rank ("combined"). Factors enter the investable universe only after publication. Panel (b) reports the time-series correlation between historical t-statistics and monotonicity across published factors (black line, right axis), together with the number of overlapping factors selected by the two ranking rules and the size of the eligible factor pool each month (bars, left axis). Panel (c) reports the same four strategies' post-publication risk and performance: mean return is the HAC (Newey-West) monthly mean of the strategy's self-financing spread return, in bps/month; CAPM, FF3, FF5+Mom, and q-factor alphas are from time-series regressions of the same spread return on each model's benchmark factors, also in bps/month (q-factor is the Hou-Xue-Zhang 4-factor model, R_{MKT}, R_{ME}, R_{IA}, R_{ROE}); t-values in parentheses below the mean return and each alpha are HAC/Newey-West; Sharpe is annualized; certainty-equivalent (CE) return is the mean-variance approximation $\text{CE} = 12\mu - 6\gamma\sigma^2$ (monthly μ, σ), annualized, at risk-aversion $\gamma \in \{1, 3, 5, 10\}$. Sample: January 2000 through December 2024 (300 months), monthly U.S. stock return data from CRSP and accounting data from Compustat.

(a) Cumulative post-publication return



(b) t-stat / monotonicity correlation and overlap



(c) Post-publication risk and performance

	Return / Alpha (bps/mo)						CE Return (ann., %)			
Strategy	Mean	CAPM	FF3	FF5+Mom	q-factor	Sharpe	$\gamma = 1$	$\gamma = 3$	$\gamma = 5$	$\gamma = 10$
All-factor	28.28***	28.7***	27.8***	23.5***	24.4**	0.67	3.27	3.01	2.76	2.12
	(3.21)	(3.36)	(3.53)	(2.71)	(2.28)					
Top 30% t-stat	29.41***	31.0***	30.1***	29.5***	30.6***	0.77	3.43	3.22	3.01	2.49
	(3.60)	(3.84)	(4.05)	(3.62)	(3.26)					
Top 30% monotonicity	34.30**	34.3**	32.9**	29.5*	32.8*	0.50	3.78	3.12	2.45	0.79
	(2.35)	(2.42)	(2.46)	(1.93)	(1.85)					
Top 30% combined	39.94***	40.7***	39.6***	37.2***	39.7***	0.71	4.57	4.11	3.66	2.52
	(3.21)	(3.35)	(3.47)	(2.90)	(2.64)					

C. Long/Short Leg Decomposition

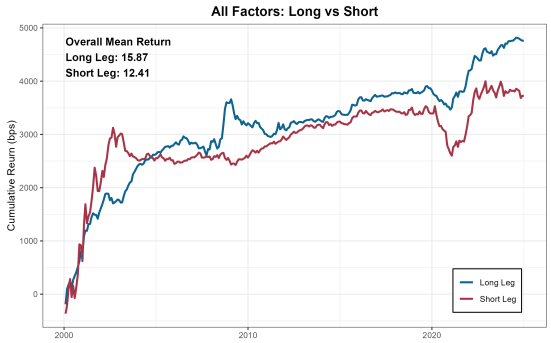
To decompose each strategy's post-publication performance into its long and short legs, Figure 4 plots cumulative long-leg and short-leg returns separately for all four strategies (panels (a)–(d)), together with a summary of mean return, Sharpe ratio, and certainty-

equivalent return by leg (panel (e)). The four strategies' long/short structure differs sharply. All-factor's two legs are close in magnitude and both individually significant (long 15.87 bps/mo, $t = 3.34$; short 12.41 bps/mo, $t = 2.02$). The t-statistic strategy's legs are balanced but individually *insignificant* (long 13.13 bps, $t = 1.45$; short 16.28 bps, $t = 1.36$) — its combined-spread significance (Table II) comes from the two legs adding constructively, not from either leg alone. Monotonicity's long leg does essentially all the work (26.62 bps, $t = 4.20$) while its short leg is statistically indistinguishable from zero (7.68 bps, $t = 0.66$). The combined strategy pushes this pattern furthest: its long leg (34.12 bps, $t = 4.15$) is the best-performing leg of any strategy in this table, while its short leg (5.82 bps, $t = 0.43$) is the worst — combining the two selection criteria concentrates the long-side signal further rather than diversifying the short-side weakness away. Panel (e) shows the same pattern in certainty-equivalent terms: every long leg's CE stays positive through $\gamma = 10$, while three of the four short legs' CE turns negative by $\gamma = 10$ (monotonicity -1.00%, combined -2.46%, t-statistic -0.63%) — a risk-averse investor holding only the short leg of any strategy but t-statistic's own would be worse off than holding cash.

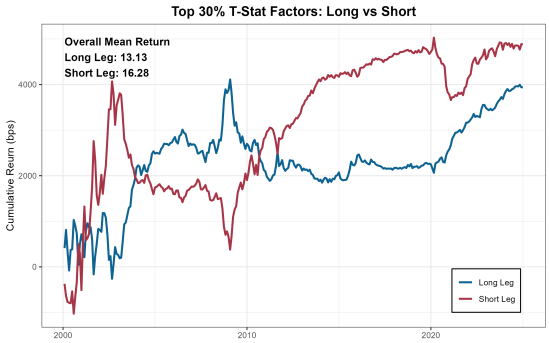
This asymmetry is not anomalous but consistent with the broader long/short factor-decomposition literature. Israel and Moskowitz (2013) find that long positions account for nearly all of the historical size premium, roughly 60% of value, and about half of momentum, establishing that a factor's economic content residing predominantly in its long leg is the norm rather than the exception. Blitz et al. (2020) reach a similar conclusion in a beta-hedged setting: across five equity factors, combined long legs diversify each other better and achieve a higher Sharpe ratio than combined short legs, and long-leg alpha survives controlling for the short legs while the reverse does not hold. My results echo this pattern closely — monotonicity's and the combined strategy's post-publication performance are effectively long-leg phenomena. If a strategy's short leg contributes little, or actively detracts once risk is priced in, a long-only implementation is not merely a fallback for investors unable or unwilling to short — it can be the more appealing implementation outright, retaining nearly all of the long–short spread's performance while shedding a leg that adds risk without adding return.

Figure III
Long/Short Leg Decomposition, All Four Strategies

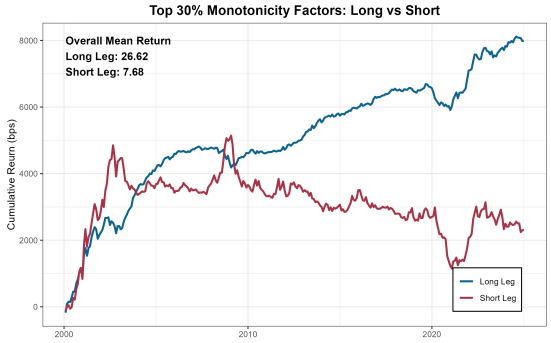
Panels (a)–(d) plot cumulative post-publication long-leg and short-leg returns (bps) from 2000 onward for the All-factor, t-statistic, monotonicity, and combined strategies, respectively; annotated means are the full-sample average monthly leg return. Panel (e) reports, for each of the eight strategy-legs: the HAC (Newey-West) mean monthly return (bps/mo, with t-value in parentheses below), the annualized Sharpe ratio, and the certainty-equivalent annual return ($\text{CE} = 12\mu - 6\gamma\sigma^2$) at risk-aversion $\gamma \in \{1, 3, 5, 10\}$. Significance stars (*, **, *** at the 10%, 5%, 1% levels) are based on the HAC t-value. Sample: January 2000 through December 2024.



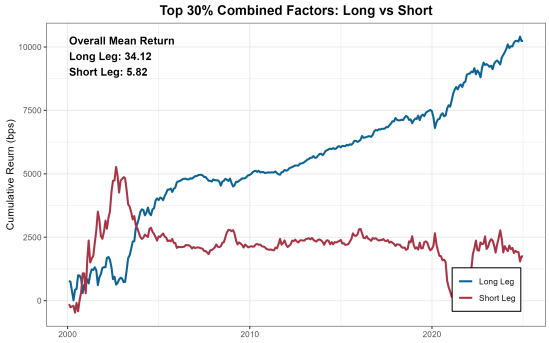
(a) All-factor



(b) Top 30% t-statistic



(c) Top 30% monotonicity



(d) Top 30% combined

	Mean (bps/mo)	Sharpe	CE γ=1	CE γ=3	CE γ=5	CE γ=10
All-factor – Long	15.87***	0.85	1.88	1.83	1.78	1.66
	(3.34)					
All-factor – Short	12.41**	0.41	1.42	1.29	1.15	0.82
	(2.02)					
Top 30% t-stat – Long	13.13	0.28	1.42	1.10	0.78	−0.01
	(1.45)					
Top 30% t-stat – Short	16.28	0.27	1.70	1.18	0.66	−0.63
	(1.36)					
Top 30% monotonicity – Long	26.62***	0.97	3.14	3.03	2.93	2.66
	(4.20)					
Top 30% monotonicity – Short	7.68	0.15	0.73	0.35	−0.04	−1.00
	(0.66)					
Top 30% combined – Long	34.12***	0.81	3.97	3.71	3.46	2.82
	(4.15)					
Top 30% combined – Short	5.82	0.09	0.38	−0.25	−0.88	−2.46
	(0.43)					

(e) Long/short leg summary

D. Trading Costs and Frictions

To assess whether each strategy is likely to survive trading costs and implementation frictions, I first examine two cost *proxies* — the market equity of the stocks each strategy trades, and each strategy's own turnover — before turning to the two priced cost models below. Figure 6(a) plots the average market equity of traded stocks for all four strategies. Monotonicity trades the largest stocks of any strategy on average ($6.32 million thousand, i.e. $6.32 billion), ahead of All-factor ($5.97B), the combined strategy ($5.67B), and t-statistic ($4.62B, the smallest of the four).

Figure 6(b) reports each strategy's own cumulative turnover, using the same average single-factor (non-netted) convention as Table III: each month's value is the equal-weighted average, across the currently-selected factors, of that factor's own turnover — as if each factor traded in its own separate account. Here the picture is less favorable to monotonicity than Figure 6(a) alone would suggest. All-factor has the lowest average turnover (0.38/month), but the three selective strategies cluster much closer together: monotonicity (0.476) and t-statistic (0.475) are essentially identical, and the combined strategy has the *highest* turnover of all four (0.51) — higher than either individual signal on its own. Monotonicity's larger average stock size does not translate into a materially lower turnover profile than t-statistic's; whatever cost advantage monotonicity has (Table III, Panel A: net Sharpe 0.13–0.17 versus t-statistic's 0.11–0.19) comes almost entirely from trading larger, cheaper-to-cross names, not from trading less often.

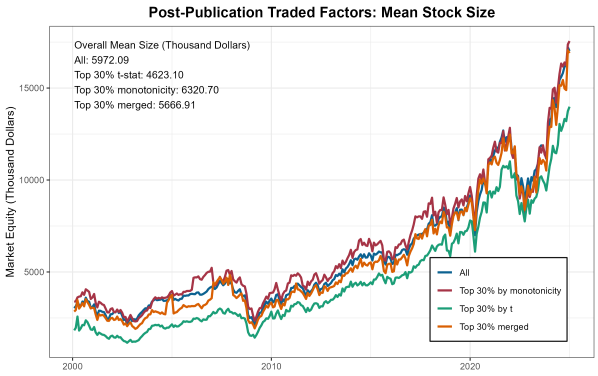
Market equity and turnover are useful cost *proxies*, but they do not answer the economic question directly: what does each strategy's edge look like once trading costs are actually priced and subtracted? I apply two complementary cost models to each factor's own trading activity, following the plan's stated convention of using conservative, transparent assumptions rather than fitted or omitted costs. Both models start from the same underlying trade: each currently-selected characteristic's own top-minus-bottom quintile rebalancing, priced using its own turnover and its own extreme-leg stocks' characteristics (bid-ask spread, borrow-fee tier, realized volatility, dollar volume) — not a netted portfolio-level position. I report each strategy's cost as the equal-weighted average, each month, of its currently-selected factors' own standalone cost — i.e., as if every factor traded in its own separate account, with no credit for offsetting long/short positions in the same stock across different factors that happen to be selected simultaneously. This "naive gross" convention isolates the cost properties of the *average single factor* a strategy selects, rather than the netting benefit a diversified book can capture.

The **simple-schedule model** prices each factor's own half effective-bid-ask-spread cost of turnover (a 21-day Corwin–Schultz estimator) plus a short-borrow financing cost on its short leg (three illustrative fee schedules, from a uniform 25 bps/year to a size-tiered schedule

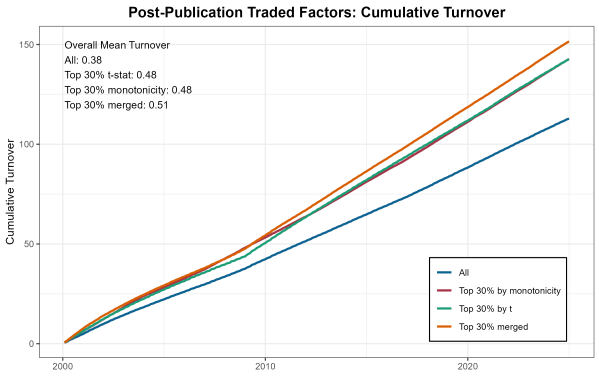
Figure IV

Avg. Stock Market Equity and Cumulative Turnover of Traded Factors

Panel (a) plots, from 2000 onward, the time series of the average market equity (thousands of dollars) of traded stocks under all four strategies. Panel (b) plots each strategy's cumulative turnover, where each month's turnover is the equal-weighted average, across the currently-selected factors, of that factor's own turnover (average single-factor/non-netted, the same convention as Table III) — not the sum of absolute changes in a netted cross-factor position. Factors enter the investable universe only after publication. Sample: January 2000 through December 2024.



(a) Avg. market equity of traded stocks



(b) Cumulative turnover (average single-factor)

reaching 800 bps/year for the smallest names); it is scale-invariant, since both cost components are already expressed as a fraction of the factor's own notional and do not depend on how much capital is deployed. The **impact model** instead prices each factor's trade using the square-root law of market impact (cost per dollar traded rises with the square root of the trade's size relative to the stock's own trading volume): at a given assumed strategy AUM, each currently-selected factor is assumed to receive an equal $1/N_t$ share of that AUM (N_t = the number of factors selected that month), matching the equal-weighted portfolio return construction used throughout this paper — so a factor's dollar trade size, and hence its impact cost, depends on how many *other* factors are selected that month, not on its own turnover alone. Both models use the same $k = 1$ (Grinold–Kahn-style (Grinold and Kahn, 2000), illustrative, not fitted) impact-law constant and the same three borrow-fee schedules used throughout this analysis.

Table III tells a more cautious story than the market-equity/ turnover proxies above. Panel A shows that once the simple-schedule cost is applied, the t-statistic strategy's best-gross-Sharpe ranking inverts: it has the highest gross Sharpe (0.77) but the lowest net Sharpe under every borrow schedule (0.11–0.19), because its higher turnover consumes more of its edge; the combined strategy holds up best (0.23–0.29), followed by All-factor (0.18–0.25). Monotonicity's net Sharpe (0.13–0.17) sits closest to t-statistic's despite a much larger gross-Sharpe gap between the two.

Panel B shows the same inversion, sharply amplified by scale. Because each selected factor

Table III: Trading Costs, Naive (Non-Netted) Single-Factor Average, All Four Strategies

Table III reports, for each strategy, the equal-weighted average across its currently-selected factors of that factor's own gross return and cost, month by month — naive costing, no cross-factor netting. Panel A: simple-schedule model, three borrow-fee schedules (Low = uniform 25 bps/yr; Tiered = 15–300 bps/yr by size; D'Avolio (D'Avolio, 2002) = 15–800 bps/yr, specials-heavy). Panel B: impact model, four illustrative AUM levels, $k = 1$. All figures are mean monthly return in bps, with the annualized Sharpe ratio in parentheses. Sample: January 2000 through December 2024.

Panel A: Simple-Schedule Model

Strategy	Gross	Net: Low	Net: Tiered	Net: D'Avolio
All-factor	28.3 (0.67)	10.4 (0.25)	9.6 (0.23)	7.4 (0.18)
Top 30% t-stat	29.4 (0.77)	7.3 (0.19)	6.6 (0.17)	4.3 (0.11)
Top 30% monotonicity	34.3 (0.50)	11.7 (0.17)	10.8 (0.16)	8.5 (0.13)
Top 30% combined	39.9 (0.71)	16.1 (0.29)	15.2 (0.27)	12.7 (0.23)

Panel B: Impact Model

Strategy	Gross	Net: \$1M	Net: \$10M	Net: \$100M	Net: \$1B
All-factor	28.3 (0.67)	27.4 (0.65)	25.5 (0.61)	19.6 (0.47)	0.9 (0.02)
Top 30% t-stat	29.4 (0.77)	27.2 (0.72)	22.6 (0.59)	7.7 (0.20)	−39.2 (−0.90)
Top 30% monotonicity	34.3 (0.50)	32.2 (0.47)	27.5 (0.41)	12.9 (0.19)	−33.4 (−0.49)
Top 30% combined	39.9 (0.71)	37.6 (0.67)	32.6 (0.58)	16.8 (0.30)	−33.1 (−0.57)

is assumed to receive a full $1/N_t$ share of assumed AUM under naive costing, the three selective strategies — which hold far fewer factors at once ($N_t \approx 31$ for t-statistic/monotonicity/combined, versus ≈ 101 for All-factor) — concentrate a given AUM into larger per-factor trades, and impact cost rises with the square root of trade size. At the smallest AUM shown (\$1M), all four strategies still retain a healthy net Sharpe (0.47–0.72), and the ranking is not yet inverted. The gap widens quickly with scale, though: by \$100M, the three selective strategies have already fallen to 0.19–0.30 while All-factor is still at 0.47, and by \$1B, All-factor is the *only* strategy with a positive net Sharpe at all (0.02), while t-statistic, monotonicity, and combined are all sharply negative (−0.90, −0.49, −0.57).

III. Factor Dynamics

The results so far establish that both signals predict future returns and that their portfolio-level and leg-level behavior differs sharply, but they do not by themselves say *why*. This subsection takes a different angle: rather than asking how well each strategy performs, I ask which characteristics each strategy actually trades over the post-publication sample, and how the three selective strategies' rosters compare by economic family.

Table IV reports, for each of the three selective strategies and each of the six characteristic families (the same taxonomy used for the family fixed effects in Table II), how many of that family's characteristics the strategy ever trades (*Traded*) and how many of those also

rank in the strategy's own top 20 by selection frequency (*Top20 Traded*), each as a count and as a % of the family's total size; the full per-characteristic rosters underlying these counts are omitted here for space. The three strategies concentrate in different economic themes. The t-statistic strategy trades 22 of 31 INVESTMENT characteristics (71%), of which 14 (45% of the family) rank in its top 20 — including asset-growth and overinvestment characteristics such as `capex_abn`, `capx_gr1`, and `noa_at` — a cluster for which q-theory (high investment mechanically signals a low discount rate, with no mispricing required) is a long-standing, fully-rational account. Monotonicity trades TRADING FRICTIONS most broadly (20 of 29, 69%, of which 9 or 31% rank in its top 20: `beta_60m`, `betabab_1260d`, `betadown_252d`, `bidaskhl_21d`, four idiosyncratic-volatility variants, and `market_equity`) and VALUE most concentratedly at the top tier (10 of 17, 59%, of which 6 or 35% rank in its top 20). Unlike the investment cluster, there is no comparably credible fully-rational account for why idiosyncratic-volatility, beta, spread, and size characteristics should be systematically mispriced in the direction monotonicity flags — these are precisely the constructs the limits-to-arbitrage literature uses as its own proxies. The two strategies' top-20 rosters overlap in only 2 of 20 characteristics (`mispricing_mgmt` and `mispricing_perf`, Stambaugh-Yuan's own explicit mispricing composites, appearing as top-tier picks under *both* criteria).

The combined strategy's top 20 is not simply the union of its two parents: only 5 of its 20 members also appear in the t-statistic strategy's top 20, and only 8 also appear in monotonicity's (with the two mispricing composites counted in both overlaps). Profitability is the plurality family in the combined roster (7 of 20, versus 3 of 20 for t-statistic and 1 of 20 for monotonicity) — nine characteristics, including `f_score`, `ni_me`, `oaccruals_at`, `ocf_me`, `op_at`, `qmj`, and `qmj_prof`, rank in the combined top 20 despite appearing in *neither* parent's own top 20 — not extreme enough on either individual criterion to make that criterion's own top-tier cut, but solid on both simultaneously, which is exactly the population the rank-sum construction is designed to surface. The combined roster is also markedly less concentrated across families than either parent's: its largest single-family share is 35% (Profitability), versus 70% for the t-statistic strategy (INVESTMENT) and 45% for monotonicity (TRADING FRICTIONS). This is concrete evidence that combining the two measures has genuine complementary value, rather than just averaging away each measure's individual weaknesses: the combination does not simply blend two concentrated rosters into a third concentrated roster, it surfaces a distinct, more broadly diversified set of characteristics that neither measure would select on its own.

Table IV: Family Composition of Traded Factors by Strategy

Table IV reports, for each of the three selective strategies and each of the six characteristic families, two counts: *Traded*, the number of that family's characteristics the strategy ever selects into its top-30% set in at least one eligible month; and *Top20 Traded*, how many of those also rank among the strategy's own top-20 most-frequently-selected characteristics (every member of every top 20 reaches a selection frequency of exactly 1.000, so ties place more than 20 characteristics on the boundary — see the discussion above). Each cell also reports, in parentheses, that count as a share of the family's total size (all eligible characteristics belonging to that family, not strategy-specific): INTANGIBLES 31, INVESTMENT 31, Momentum 11, Profitability 30, TRADING FRICTIONS 29, VALUE 17 (149 total). Family is the same six-category taxonomy defined in Section I.B, also used for the family fixed effects in Table II. Sample: January 2000 through December 2024.

	Top 30% t-stat		Top 30% monotonicity		Combined	
Family	Traded	Top20 Traded	Traded	Top20 Traded	Traded	Top20 Traded
INTANGIBLES	10 (32%)	1 (3%)	6 (19%)	2 (6%)	7 (23%)	0 (0%)
INVESTMENT	22 (71%)	14 (45%)	2 (6%)	2 (6%)	13 (42%)	5 (16%)
Momentum	7 (64%)	1 (9%)	8 (73%)	0 (0%)	8 (73%)	1 (9%)
Profitability	12 (40%)	3 (10%)	3 (10%)	1 (3%)	12 (40%)	7 (23%)
TRADING FRICTIONS	3 (10%)	0 (0%)	20 (69%)	9 (31%)	12 (41%)	3 (10%)
VALUE	6 (35%)	1 (6%)	10 (59%)	6 (35%)	11 (65%)	4 (24%)

IV. Conclusion

This paper combines t-statistics and monotonicity together to show that either diagnostic used alone leaves real predictive power on the table. Both measures predict out-of-sample returns and alphas on their own, and continue to do so essentially unchanged when entered jointly, consistent with their near-zero in-sample correlation: each carries information the other does not. But neither survives the same scrutiny equally well. Once the comparison is tightened to weigh factors only against others in their own economic family, the two diverge: monotonicity's coefficient survives while the t-statistic's does not, indicating that the t-statistic's advantage operates mainly across families rather than within them. Replacing the two separate predictors with their summed rank outperforms either alone under this same demanding specification, achieving the highest R^2 of any specification in the paper despite collapsing two variables into one—the clearest cross-sectional evidence that a single in-sample diagnostic, however carefully constructed, is an incomplete description of a factor's quality.

This complementarity is not just a cross-sectional curiosity; it shows up in tradable portfolio performance. The t-statistic strategy has the highest gross Sharpe ratio, but the combined strategy dominates on a certainty-equivalent (CE) basis at every level of risk aversion considered, including for a highly risk-averse investor. That ranking inverts again once trading costs and market impact are priced in: the t-statistic strategy's higher turnover consumes enough of its edge that its best-gross-Sharpe ranking becomes the worst net-of-cost Sharpe of the four strategies, while the combined strategy holds up best net of costs. Monotonicity

alone fares differently still: it is the highest-mean-return, highest-volatility strategy of the four, and its edge—like the combined strategy's—is concentrated almost entirely in the long leg, with a short leg that is statistically indistinguishable from zero and value-destroying once risk aversion is priced in. No single measure, and no single strategy built from one, dominates on every dimension an investor might care about; only the combination does, and even then only on some of them.

Examining which characteristics each strategy actually trades points to why combining works. The t-statistic strategy concentrates in INVESTMENT characteristics, a cluster with a long-standing, fully-rational q-theory account; monotonicity concentrates instead in TRADING FRICTIONS and VALUE characteristics more closely associated with limits-to-arbitrage proxies than with a rational risk story. The two strategies' top-20 rosters overlap in only 2 of 20 characteristics. The combined strategy's roster is not simply their union: most of its members appear in neither parent's own top tier, and its family composition is markedly less concentrated than either parent's alone. Combining the two measures does not average away each one's individual weaknesses; it surfaces a distinct population of characteristics that neither single diagnostic would flag as strong on its own—direct evidence that the enhanced predictability documented throughout this paper comes from genuine complementary information, not from noise cancellation.

Several questions remain open. The baseline monotonicity measure classifies a month as monotonic only when quintile returns are strictly non-decreasing; a more permissive measure that tolerates small reversals, or weights the *degree* of monotonicity rather than treating it as binary, may capture additional structure that the current measure understates. And while this paper's evidence is consistent with monotonicity capturing information the t-statistic does not, distinguishing whether that information ultimately reflects compensation for risk, correction of mispricing, or some combination of the two remains an open question for future work. What this paper does establish is narrower but, I think, more robust: a single in-sample diagnostic—however carefully constructed—is not enough to separate factors that will keep working out of sample from those that will not, and combining two near-orthogonal diagnostics recovers predictive power that neither captures alone.

References

Blitz, D., Baltussen, G., and van Vliet, P. (2020). When equity factors drop their shorts. *Financial Analysts Journal*, 76(4):73–99.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264.

D'Avolio, G. (2002). The market for borrowing stock. *Journal of Financial Economics*, 66(2-3):271–306.

Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk*. McGraw-Hill, New York, 2nd edition.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133.

Israel, R. and Moskowitz, T. J. (2013). The role of shorting, firm size, and time on market anomalies. *Journal of Financial Economics*, 108(2):275–301.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32.

Patton, A. J. and Timmermann, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98(3):605–625.

Shumway, T. (1997). The delisting bias in CRSP data. *Journal of Finance*, 52(1):327–340. NOT in the original manuscript's reference list – added here because Section I cites Shumway (1997) for the delisting-return convention with no corresponding entry. Verify before submission.