JOIM
www.joim.com

JOURNAL OF INVESTMENT MANAGEMENT, Vol. 19, No. 4, (2021), pp. 15–57

HOW DO FACTOR PREMIA VARY OVER TIME? A CENTURY OF EVIDENCE

Antti Ilmanen[a], *Ronen Israel*[a], *Rachel Lee*[a],
Tobias J. Moskowitz[a,b,c,*] *and Ashwin Thapar*[a]

Evaluating how factor premia vary over time and across asset classes is challenging due to limited time series data, especially outside of US equities. We examine four prominent factors across six asset classes over a century. We find little evidence for arbitrage activity influencing returns, though some novel evidence of overfitting biases. We identify meaningful time variation in risk-adjusted factor returns that appears unrelated to macroeconomic risks, supporting other theories of dynamic return premia. Attempting to capture this variation, we evaluate various factor-timing strategies, but find relatively modest predictability that likely fails to overcome implementation costs.

An abundance of empirical research focuses on unconditional factor premia that capture cross-sectional differences in asset prices. However, much less is known about how these premia vary over time, despite a large literature in asset pricing theory advocating for dynamic models.[1] Identifying and testing models of conditional return premia are especially challenging due to limited time series data used in the previous studies, and a tendency to narrowly focus on just US equities. We study how much variation in factor premia exists using unique data from almost a century across six different asset classes. The longer and broader sample offers advantages in identifying conditional expected returns both statistically and economically, providing more powerful tests to detect changes in factor premia.

We study four long-short factors that capture cross-sectional variation in returns in multiple asset classes, are measurable over our century of data, and are most common to empirical asset pricing models: value, momentum, carry, and defensive. While the literature has produced a proliferation of hundreds of factors, mostly to explain the cross-section of US equity returns, many have been questioned due

[a]AQR Capital
[b]Yale University, E-mail: tobias.moskowitz@yale.edu
[c]NBER
[*]Corresponding author.

to meager statistical support and lack of robust out-of-sample evidence (Harvey *et al*., 2016; McLean and Pontiff, 2016; Hou *et al*., 2020).[2] The handful of prominent factors that we focus on have strong in- and out-of-sample support in many markets and are commonly employed by quantitative investors (Asness *et al*., 2015).[3] The robustness of their unconditional return premia is, in part, why these factors remain at the center of most academic and practitioner asset pricing research. Our additional evidence from 50 more years of historical data further boosts that claim. However, much less is known about factor premia variation over time, which we evaluate in this study using our extensive data set.

First, we find significant evidence of variation in risk-adjusted returns, driven by highly time-varying risk, in spite of returns remaining relatively constant. Second, we seek to understand the economic drivers of this variation. We first rule out spurious or non-economic sources of variation, such as overfitting in the original studies, providing more conviction in the likelihood of our identified factors persisting out of sample. Specifically, with an additional five decades of out-of-sample evidence spanning six asset classes, we employ a novel test of overfitting bias by splitting the sample for each factor into three subperiods: the original sample period in which the factor was discovered, the period before the original sample period's start date, and the post-publication period after the factor's discovery. This partition separates the influence of data mining from another potential source of time variation in factor premia—informed trading by arbitrageurs (Schwert, 2003; McLean and Pontiff, 2016). This latter effect also has important investment implications which can lead to disappointing performance and higher risk and trading costs associated with these strategies (Alquist *et al*., 2019). We find evidence consistent with overfitting in the original sample, but little evidence that arbitrage activity has altered factor premia. Moreover, significant variation in factor premia remains that is unrelated to data mining or arbitrage activity.

To investigate the additional source of variation in factor premia, we attempt to link factor variation to economic sources by appealing to a variety of models that generate expected return dynamics. The long time series across many asset markets provides a greater opportunity to measure economic shocks and events. Conducting a broad search for economic exposure, drawing inspiration from a variety of asset pricing theories, we examine macroeconomic variables related to business cycles, growth, and monetary policy (Breeden, 1979; Campbell and Cochrane, 1999; Lettau and Ludvigson, 2001, 2009; Bansal and Yaron, 2004; Greenwald *et al*., 2014; Tsai and Wachter, 2015; Gabaix, 2012). Chen *et al*. (1986) show that some of these variables capture variation in stock returns. We also examine political uncertainty (Baker *et al*., 2016; Caldara and Iocoviello, 2018), volatility risk, downside risk, tail risk, and crash risk (Brunnermeier *et al*., 2008; Lettau *et al*., 2014; Jiang and Kelly, 2014), liquidity risk (Pastor and Stambaugh, 2003; Acharya and Pedersen, 2005), and investor sentiment (Baker and Wurgler, 2006). The additional 50 years of economic events and the breadth of assets improve the power of our tests. However, we fail to find significant exposures of the factor returns to economic activity or news, although we find some variation in the risk and correlation structure of the factors to the economic environment. This evidence challenges many macro-asset pricing models attempting to explain these factor returns (Gomes *et al*., 2003; Carlson *et al*., 2004; Zhang, 2005; Hou *et al*., 2015; Lettau and Wachter, 2007; Gormsen and Lazarus, 2019) and suggests that other sources of variation (e.g., behavioral models) may be driving the dynamics in factor risk-adjusted returns.

Finally, to answer how much of the time variation in risk-adjusted factor returns can be captured in practice, we examine a host of factor-timing approaches proposed in the literature. While the literature predominantly focuses on applying these approaches exclusively to equity factors, we address whether the same conditional information can also forecast factor returns in currencies, bonds, and commodities.

We take a broad approach, studying nearly a dozen timing signals using multiple methodologies applied to all six asset classes and four factors. We focus on returns to implementable investment strategies, which has the advantage of making different models and methods easy to compare and provides economic magnitudes. We assess the marginal benefit of factor timing to a diversified static factor portfolio to quantify the economic impact of timing to a practical factor investment model.

The result of this comprehensive search, however, yields fairly weak and inconsistent evidence for factor timing. The most consistent results come from using valuation spreads to time factors, which supports other findings in the literature for equity factors (Asness *et al.*, 2000; Cohen *et al.*, 2003), and is consistent with theory. We also find that imposing economic restrictions from theory on timing models (e.g., Campbell and Thompson, 2007) further improves out-of-sample predictability. However, even for the best performing factor timing strategies that are statistically significant, their economic impact is modest, especially after accounting for real-world implementation costs.

The rest of this paper is organized as follows. Section 1 describes our unique data and factor construction, and provides summary statistics of the factors over time. Section 2 partitions the sample into pre-, original, and post-sample periods to test for the influence of data mining biases and arbitrage activity on factor return variation. Section 3 attempts to relate time variation in the factors to economic sources motivated by theory. Section 4 analyzes conditional factor return premia using timing models with an array of conditioning information and methods. Section 5 concludes.

1 Data, Factor Construction, and Summary Statistics

We describe our data, factor construction, and present summary statistics over the last century.[4]

1.1 Data

We collect monthly asset returns and economic fundamental data going back as far as February 1877, though we start our series in 1926 to ensure broad enough asset class coverage. Our main data source is Global Financial Data, supplemented by Bloomberg and DataStream. The data cover equity indices, government bonds, currencies, and commodities. We also examine nearly a century of returns on individual stocks in the US from the Center for Research in Security Prices (CRSP), and add individual stock return data from 21 international markets beginning in 1984 from Worldscope. The individual stock universes comprise 90% of the total market capitalization of each market, and hence exclude the smallest stocks. We report backtest results starting in 1926, when CRSP data is first available for individual stocks. Appendix A provides a detailed description of our data and their sources. Relative to typical studies that focus solely on US stocks, our analysis benefits from information from a broader cross-section of assets and securities, and uses a much longer time series outside of stocks than has been previously examined. Our sample contains equity indices from 23 countries, government bonds from 13 nations, 20 exchange rates, and 30 commodities. Table B1 in Appendix B reports summary statistics on each asset in our sample (except for individual stocks).

1.2 Factor definitions

We construct cross-sectional value, momentum, carry, and defensive factor portfolios within each asset class, using the simplest, best-documented measures to reduce data mining concerns.

Value. We follow simple value measures used in the literature to capture "cheap" and "expensive" securities relative to fundamentals within an asset class. We make no qualifying claims whether assets are "cheap" due to risk or because they are mispriced. "Cheap" ("expensive") here simply means a low (high) price relative to some fundamental, or equivalently a high (low) expected return. For individual equities, we use the book-to-market ratio following Fama and French (1993, 2012). For global equity indices, we use the aggregate 10-year cyclically-adjusted price-to-earnings ratio CAPE (value-weighted average P/E ratio for all constituent firms in the index).[5] For global bonds, we use the 10-year real bond yield (Brooks and Moskowitz, 2018), which is the difference between nominal yields and expected inflation, using 3-year trailing changes in the Consumer Price Index as a proxy for inflation expectations.[6] For currencies, we use deviations from Purchasing Power Parity (PPP) exchange rates, compiled from the Penn World Tables, supplemented by OECD databases and reported inflation indices,[7] and for commodities we use the negative of 5-year changes in spot prices, following Asness *et al.* (2013) and motivated by DeBondt and Thaler (1985) and Fama and French (1996).

Momentum. We use a uniform measure of momentum across all asset classes: the past 12-month cumulative excess-of-cash return on an asset, following Jegadeesh and Titman (1993). Our factor portfolios skip the most recent month's return to avoid any microstructure effects, such as bid–ask bounce, that may induce negative short-term autocorrelation.[8]

Carry. We define carry as in Koijen *et al.* (2018), which is the expected return on an asset assuming market conditions are unchanged. For equity indices, carry is the futures-to-spot discount of the front month contract, where prior to 1990 when futures discount data is available, we use excess-of-cash dividend yield. For global currencies, carry is the short-term interest rate differential between the two countries (difference in 3-month LIBOR rates or closest 3-month equivalent unsecured lending rates).[9] For bonds, carry is the 10-year term spread (10-year yield minus 3-month interest rate). For commodity futures, carry is the return from holding a futures contract if there is no shift in the futures curve, measured by the percent difference in prices between the nearest and next-nearest-to-maturity contract. We do not construct a carry strategy for individual stocks because carry and value are nearly identical here and there are no futures on individual names.

Defensive. We use the (negated) beta of the asset with respect to its local market index following Frazzini and Pedersen (2013). For global equity indices and bonds, betas are estimated from a 36-month rolling regression of asset returns on the equal-weighted returns of all country indices and bonds, respectively. We do not construct a defensive strategy for currencies because there is no logical market index. We do not construct a defensive strategy for commodities because returns from different commodities do not share a common market component.

1.3 Factor portfolio construction

We form zero-cost, one dollar long and short factor portfolios for each asset class using their respective value, momentum, and carry characteristics as defined above. For defensive factors, we form constant zero-beta (one beta long, one beta short) portfolios which are not zero-cost, because the dollar notional on the long side (lower-beta) needs to be higher than the short side (higher-beta) in order to stay beta-neutral. This approach is consistent with the literature, for example, the betting-against-beta (BAB) factor of Frazzini and Pedersen (2013).

For each security i at time t with characteristic $S_{it} \in$ (value, momentum, carry, defensive) we first sort securities on the characteristic and assign weights based on each security's cross-sectionally demeaned ranks within the asset class, where the weights sum to zero. Specifically, the weight on security i at time t is

$$w_{it}^{S} = c_t(\text{rank}(S_{it}) - \Sigma_i \text{rank}(S_{it})/N), \quad (1)$$

where we include a scaling factor c_t such that the overall portfolio is scaled to a dollar-neutral long-short portfolio, except in the case of defensive, where we use a separate constant for long and short legs of the strategy so that the resulting portfolio is beta-neutral. The return on the portfolio is

$$r_t^S = \sum_i w_{it-1}^S r_{it}.$$

We also form a multifactor portfolio that combines all four factors. Specifically, we use the average rank of securities across all four characteristics (value, momentum, carry, and defensive) to provide the weights in Equation (1), for the multifactor (MF) portfolio

$$w_{it}^{MF} = c_t \left[\frac{1}{4} \sum_{k \in (\text{V,M,C,D})} \text{rank}(S_{it}^k) - \frac{1}{N} \sum_i \left(\frac{1}{4} \sum_{k \in (\text{V,M,C,D})} \text{rank}(S_{it}^k) \right) \right].$$

Finally, we combine factor portfolios across assets classes by weighting each asset class by the inverse of its standard deviation (estimated using the past 36 months of returns). The portfolio construction and weighting scheme follow Moskowitz *et al*. (2012) and Asness *et al*. (2013, 2015).

1.4 Summary statistics over a century

Table 1 reports summary statistics of the returns of the factor portfolios over the last century by asset class. The premia are all positive for each factor in each asset class and the majority are statistically significant. The Sharpe ratios for value, momentum, carry, and defensive when applied across all asset classes are 0.53, 0.64, 0.57, and 0.68, respectively. Sharpe ratios are generally of greater magnitude in stock selection than in other asset classes, due to the greater breadth in individual equities leading to better diversification benefits, or potentially the result of overfitting the factor definitions in the original US equity sample.[10] When pooling across asset classes, the t-statistics of the mean returns range from 5.1 for value to more than 6.6 for defensive, easily rejecting a null hypothesis of the factors being uncompensated. The multifactor portfolio that combines all four factors across asset classes has a Sharpe ratio of 1.46 with a t-statistic of 14.2, indicating large diversification benefits from combining different factors and different asset classes. Our broader and century-long data set, which contains meaningful out-of-sample evidence relative to original studies, shows overwhelming evidence of positive return premia for these asset pricing factors and casts serious doubt on these factors being data mined.

1.5 Time-varying factor premia

Figure 1 plots the Sharpe ratios of each factor in each asset class by decade. The premia

Table 1 Factor return premia over a century.

Asset class	Factor/Style	Mean	St.dev	Sharpe	*t-stat*	Skew	Kurt	Start date	End date
Panel A: Summary statistics on raw returns									
US stocks	Value	3.9%	14.9%	0.26	2.54	3.3	29.2	Jul-1926	Nov-2020
	Momentum	8.2%	15.8%	0.52	5.02	−3.0	25.6	Jan- 1927	Nov-2020
	Defensive	8.1%	11.2%	0.72	6.86	−0.7	7.2	Dec-1930	Nov-2020
	Multifactor	6.7%	6.1%	1.11	10.79	0.5	9.8	Jul-1926	Nov-2020
International stocks	Value	4.3%	9.7%	0.44	2.64	0.1	6.2	Jul-1984	Nov-2020
	Momentum	8.4%	12.3%	0.68	4.08	−1.1	6.1	Jan-1985	Nov-2020
	Defensive	9.5%	9.7%	0.98	5.67	0.0	1.0	Feb-1987	Nov-2020
	Multifactor	7.4%	5.4%	1.36	8.23	0.5	2.9	Jul-1984	Nov-2020
Commodities	Value	7.7%	20.0%	0.38	3.59	0.2	2.0	Jul-1926	Nov-2020
	Momentum	5.7%	20.0%	0.29	2.70	−0.4	2.4	Jul-1926	Nov-2020
	Carry	4.6%	16.8%	0.28	2.58	−0.5	2.5	Jul-1926	Nov-2020
	Multifactor	6.0%	9.4%	0.64	5.99	−0.2	2.8	Jul-1926	Nov-2020
Equity indices	Value	3.1%	14.0%	0.22	2.18	−0.3	7.5	Jul-1926	Nov-2020
	Momentum	7.0%	15.4%	0.46	4.43	0.0	4.5	Jul-1926	Nov-2020
	Carry	2.2%	12.8%	0.17	1.70	−0.5	4.1	Jul-1926	Nov-2020
	Defensive	4.2%	12.0%	0.35	3.42	−0.3	3.0	Jul-1926	Nov-2020
	Multifactor	4.2%	6.6%	0.63	6.09	−0.8	6.3	Jul-1926	Nov-2020
Fixed income	Value	1.3%	4.6%	0.29	2.84	0.1	8.0	Jul-1926	Nov-2020
	Momentum	0.6%	4.7%	0.12	1.16	−0.9	7.2	Jul-1926	Nov-2020
	Carry	2.9%	4.4%	0.65	6.33	0.3	8.4	Jul-1926	Nov-2020
	Defensive	0.0%	4.4%	0.00	0.02	−0.1	6.4	Jul-1926	Nov-2020
	Multifactor	1.2%	2.3%	0.52	5.07	−1.1	15.5	Jul-1926	Nov-2020
Currencies	Value	3.2%	5.3%	0.60	4.10	0.1	1.7	Apr-1974	Nov-2020
	Momentum	1.1%	6.7%	0.16	1.08	−0.5	0.6	Feb-1974	Nov-2020
	Carry	2.8%	6.6%	0.42	2.90	−0.6	3.4	Feb-1974	Nov-2020
	Multifactor	2.3%	3.8%	0.62	4.25	−0.4	1.9	Feb-1974	Nov-2020
All asset classes	Value	2.6%	4.9%	0.53	5.12	1.8	15.5	Jul-1926	Nov-2020
	Momentum	3.6%	5.7%	0.64	6.18	−1.4	9.1	Jul-1926	Nov-2020
	Carry	2.6%	4.5%	0.57	5.54	−0.4	3.6	Jul-1926	Nov-2020
	Defensive	3.0%	4.5%	0.68	6.57	−0.7	5.4	Jul-1926	Nov-2020
	Multifactor	3.2%	2.2%	1.46	14.17	−0.3	3.0	Jul-1926	Nov-2020
Avg. single factor, single asset		4.4%	11.1%	0.40	3.29	−0.2	6.8		
Avg. multi-factor, single asset		4.6%	5.6%	0.81	6.74	−0.2	6.5		
Avg. single factor, multi-asset		3.0%	4.9%	0.60	5.85	−0.2	8.4		

Notes: Reported are annualized means, standard deviations, Sharpe ratios, t-statistics of the mean, skewness, and kurtosis of the returns to factor portfolios in each asset class over the last century, with sample periods reported in the last column. Results are reported for US stocks, international stocks, commodities, equity index futures, bonds, and currencies separately, as well as for all asset classes combined (equal volatility-weighted average of the asset classes). The last three rows report the averages of each variable across all single factor, single asset class portfolios, across all multifactor, single asset class portfolios, and across all single factor, multi-asset portfolios.

Figure 1 Factor return premia by decade.

The figure plots the Sharpe ratios of each factor in each asset class decade-by-decade. The last two graphs plot Sharpe ratios by decade for the portfolio diversified across factors within each asset class ("multifactor") and for portfolios diversified across asset classes for each factor ("all assets"), respectively.

for each factor are positive in the majority of decades, with only a handful of instances of decade-long underperformance for a factor. The premia appear largely stable over time, with little variation across decades, a result we confirm formally below with statistical tests. The multifactor portfolios exhibit even more stable premia over decades, and multi-asset versions of the factors are also reasonably stable. Finally, highlighted in the last graph is the multifactor, multi-asset portfolio, which exhibits consistently strong performance across decades over the last century.

Our century of data has more power to detect time-varying premia than prior studies. We test formally for variation in factor premia and risks over time using (asset pricing) model-free statistical tests. Table B2 reports results from structural break tests in the mean and variance of returns of the factors by asset class. We can reject a constant mean over decades in our sample for 8 of the 20 factor-by-asset class portfolios and four of the seven multi-factor portfolios. Looking at multi-asset versions of each factor, carry and defensive each exhibit significant mean changes over decades, while value and momentum fail to show significant time variation. For comparison, running the same tests on the equity market portfolio, we cannot reject that the equity premium has had a constant mean over the last century. Hence, although the majority of the literature on detecting time-varying expected returns focuses on the equity premium, there is more robust evidence of return variation over time for the factor premia we study. The results suggest the modest presence of time variation in factor premia and our data appear to provide enough power to detect it. The implications for investment are that factor premia variation can distort return expectations over time. The question remains whether this variation is predictable and what might be driving it.

We also analyze time variation in the risk of factor portfolios. The structural break tests on the variance of these factor portfolios (Table B2, Panel B) indicate ample evidence of variance changes over time for all factors (and for the market). This, coupled with time-varying returns, implies significant time variation in factor Sharpe ratios. Not only is this variation relevant to practitioners attempting to harvest these premia, but it also offers new observations to test various dynamic asset pricing theories in an attempt to understand the source of this variation.

Finally, we turn our focus to the multi-factor portfolio that is diversified across all factors, where there is more limited evidence of mean return variation but still robust evidence of risk variance. This finding suggests the return per-unit-of-risk varies significantly over time, even for a portfolio that is diversified across factors. The diversification across factors does not completely ameliorate the changing risk of each factor over time. Figure B1 plots the time series of pairwise correlations between the factors (across all asset classes) using rolling monthly return data over the prior 10 years from 1936 to 2020. The graph suggests that correlations between the factors vary over time, which is also contributing to the variation in risk of the multifactor portfolio. However, there are no periods when correlations all become large enough to severely reduce diversification benefits and there is no apparent upward trend in correlations, contrary to what some have speculated.

2 Overfitting versus Informed Trading

One starting point for analysis of time variation in factor returns is potential degradation out-of-sample, as analyzed by (Schwert, 2003; Harvey *et al.*, 2016; McLean and Pontiff, 2016; Hou *et al.*, 2020). In addition to confirming this result, we seek to differentiate between two explanations behind time variation in factor returns: overfitting due to data mining of the original sample in which the factor was discovered and degradation over time due to arbitrage activity from informed traders post discovery of the factors. Our lengthy data set provides a richer comparison and more compelling tests to augment the findings in other studies.

McLean and Pontiff (2016), in their examination of 97 US equity anomalies, carve up each anomaly's sample period into in-sample, post-sample, and post-publication subperiods. They find a 26% decline in performance post-sample and a 58% decline post-publication in average returns. Since the post-sample and post-publication subperiods largely overlap, they attribute the 26% decline to overfitting and the additional 32% decline from publication due to informed trading. One potential concern with this study is the relatively short time periods (particularly the period between the data sample ending and the study being published).

Our century of data allows us to expand and improve upon this result in a novel way; we are able to additionally explore the "pre-sample" evidence on the efficacy of factors *before* their discovery that precedes the start of the original data sample. This analysis is particularly useful in distinguishing data mining biases from informed trading since the sample period before the original sample even begins could not plausibly be known to traders at the time, at least not widely.[11] Comparing the pre-sample evidence versus the original sample evidence therefore provides a clean test of overfitting biases,[12] with no influence from arbitrage activity, while comparing the post-sample evidence versus the original sample evidence measures both data mining and informed trading. Looking at the difference between the post-sample evidence and pre-sample evidence should therefore be an unbiased estimate of the influence of informed trading. Crucially, the length of our sample provides plenty of observations to deliver results with sufficient statistical power to draw inference. The results are important for practical investment because overfitting implies weaker out-of-sample expectations on the factor premia, while informed trading implies factor premia will vary with the costs and ease of arbitrage activity.

For each factor, we split the sample into three subperiods: the original sample period the factor was discovered, the pre-sample period containing data before the original sample period begins, and the post-period after the original study's publication. We follow common convention in the literature and define the original sample periods for each factor based on the most prominently cited papers that discovered these factors.[13] For currencies we have no pre-sample period since exchange rates were pegged under Bretton Woods prior to 1973.[14] As another out-of-sample test, we also compare the efficacy of factors in the original asset class in which they were discovered with their performance in other asset classes. Each of these factors was originally discovered in US equities, except for carry, which was discovered in currencies. For the factor subperiods in other asset classes, we use the same dates as those from the original study because most research on factors in other asset classes has been relatively recent (Asness *et al.*, 2013, 2015), and once a factor is discovered in one market, it seems reasonable to believe that it was applied to other markets and asset classes at or near the same time by practitioners.[15]

2.1 In- versus out-of-sample results

Table 2 reports annualized Sharpe ratios of each factor in each asset class over their respective pre-, original and post-sample periods. The factor returns are normalized by their standard deviations estimated over the prior 36 months to account for changing volatilities over time and compare premia on a per-unit-of-risk basis to enable comparisons across asset classes with different volatilities. Sharpe ratios also have a theoretical link to economic models (Hansen and Jagannathan, 1991).

We report the difference between the original sample period and an average of the pre- and post-sample periods, along with a t-statistic of whether the difference is statistically different from zero, to test for out-of-sample performance differences. We also report the difference between the original sample period's performance of the factor and its pre-sample performance as a test of overfitting that removes the potential influence of informed trading.

Panel A reports results for the value factor. For US stocks, value performs better in the original sample period than either the pre- or post-sample periods, consistent with overfitting biases, though the difference is not statistically significant. There

Table 2 Factor premia in original, post-publication, and pre-sample periods.

				Out of sample		Data mining		Arbitrage degradation	
Sharpe ratios	Pre	Original	Post	Original− (Pre & Post)	t-stat	Original− Pre	t-stat	Post− Pre	t-stat
Panel A: Value									
US stocks	0.22	0.37	−0.13	0.33	(1.46)	0.15	(0.59)	−0.36	(−1.42)
International stocks		0.39	0.34	0.06	(0.10)				
Commodities	0.22	0.50	0.44	0.16	(0.67)	0.28	(1.02)	0.23	(0.83)
Equity indices	0.12	0.56	0.07	0.46	(2.02)	0.43	(1.69)	−0.05	(−0.21)
Fixed income	0.26	0.59	0.08	0.41	(1.82)	0.33	(1.29)	−0.18	(−0.71)
Currencies		0.72	0.69	0.03	(0.09)				
All asset classes	0.38	1.00	0.42	0.59	(2.62)	0.61	(2.40)	0.04	(0.16)
Average, (p-value of F-test)				0.29	(*0.00*)	0.36	(*0.00*)	−0.06	(*0.76*)
Panel B: Momentum									
US stocks	0.84	0.93	0.68	0.16	(0.71)	0.08	(0.33)	−0.16	(−0.64)
International stocks		0.60	1.02	−0.42	(−0.59)				
Commodities	0.32	0.55	0.19	0.30	(1.26)	0.23	(0.81)	−0.13	(−0.49)
Equity indices	0.33	0.79	0.55	0.35	(1.51)	0.45	(1.75)	0.22	(0.87)
Fixed income	0.16	0.20	0.19	0.03	(0.11)	0.04	(0.16)	0.03	(0.13)
Currencies		0.40	0.08	0.32	(0.97)				
All asset classes	0.71	1.04	0.78	0.30	(1.29)	0.33	(1.27)	0.06	(0.26)
Average, (p-value of F-test)				0.15	(*0.09*)	0.23	(*0.10*)	0.00	(*0.70*)

Table 2 (*Continued*)

Sharpe ratios	Pre	Original	Post	Out of sample: Original− (Pre & Post)	Out of sample: *t*-stat	Data mining: Original− Pre	Data mining: *t*-stat	Arbitrage degradation: Post− Pre	Arbitrage degradation: *t*-stat
Panel C: Carry									
Commodities	0.21	0.40	0.58	0.00	(−0.00)	0.19	(0.52)	0.37	(1.56)
Equity indices	0.29	0.22	0.16	−0.01	(−0.03)	-0.07	(−0.20)	−0.13	(−0.59)
Fixed income	0.96	1.26	0.37	0.58	(1.65)	0.30	(0.83)	−0.60	(−2.70)
Currencies		0.33	0.51	−0.17	(−0.36)				
All asset classes	0.72	1.11	0.73	0.38	(1.09)	0.39	(1.06)	0.02	(0.08)
Average, (p-value of F-test)				0.16	(*0.24*)	0.20	(*0.39*)	−0.09	(*0.31*)
Panel D: Defensive									
US stocks	0.77	0.95	1.14	0.07	(0.34)	0.19	(0.78)	0.37	(1.03)
International stocks		0.81	1.64	−0.83	(−2.21)				
Equity indices	0.42	0.22	0.54	−0.23	(−1.09)	-0.20	(−0.87)	0.12	(0.33)
Fixed income	0.30	-0.26	0.12	−0.52	(−2.47)	-0.56	(−2.46)	−0.18	(−0.51)
All asset classes	0.62	0.62	1.51	−0.25	(−1.17)	-0.01	(−0.02)	0.88	(2.51)
Average, (p-value of F-test)				−0.35	(*0.03*)	-0.15	(*0.58*)	0.30	(*0.00*)
Panel E: Mulitfactor									
US stocks	1.41	1.83	0.78	0.75	(3.37)	0.42	(1.64)	−0.63	(−2.44)
International stocks		0.41	1.72	−1.31	(−2.34)				
Commodities	0.24	1.02	0.77	0.46	(2.04)	0.78	(2.77)	0.53	(1.87)
Equity indices	0.56	0.97	0.48	0.45	(2.02)	0.41	(1.60)	−0.08	(−0.31)
Fixed income	0.81	0.75	0.42	0.12	(0.55)	-0.06	(−0.25)	−0.39	(−1.52)
Currencies		0.97	0.63	0.34	(1.06)				
All asset classes	1.27	2.05	1.67	0.59	(2.68)	0.79	(3.10)	0.41	(1.59)
Average, (p-value of F-test)				0.20	(*0.00*)	0.47	(*0.00*)	−0.03	(*0.83*)

Notes: Reported are the annualized Sharpe ratios of each factor in each asset class over their respective pre-, original, and post-sample periods, defined using the dates in McLean and Pontiff (2016) for US equities but applied to our century of data in other asset classes. The factor returns used were normalized by the standard deviation of prior 36 months of returns to account for changing volatilities over time. For carry strategies we use the original sample dates from Meese and Rogoff (1983) and Fama (1984). We also aggregate across all asset classes into a diversified factor. We report the difference in Sharpe ratios for the original-sample versus out-of-sample periods (pre- and post-samples) along with t-statistics on the difference, the difference between the original sample and the pre-sample period as a measure of data mining and their t-statistics, and the difference between the pre- versus post-publication sample periods as a measure of arbitrage degradation and their t-statistics. Results are reported separately for value (Panel A), momentum (Panel B), carry (Panel C), defensive (Panel D), and multifactor (Panel E). Formal F-tests of the joint significance of these differences across asset classes are reported at the bottom of each panel.

is still a sizeable positive Sharpe ratio for US stock value in the pre-sample period, indicating that the value premium is not solely driven by pure spurious data mining, though the Sharpe ratio is negative in the post-sample period. Looking at the other asset classes, the out-of-sample periods for international stocks and currencies essentially match the performance of the original sample (there is no pre-sample period for these two asset classes). For the remaining asset classes, the original sample period also outperforms the out-of-sample periods. Across all asset classes, a formal test of whether the Sharpe ratio to value is the same in the original sample period versus the out-of-sample periods is rejected (p-value of 0.00), with a 49% decline in Sharpe ratio on average. This decline is even larger than what McLean and Pontiff (2016) find for the equity value factor over a much shorter sample period.

More interestingly, we turn next to the difference between the original sample period and pre-sample period: our novel and cleaner test of data mining bias since it excludes any impact from arbitrage activity. We see degradation for all asset classes, rejecting the F-test of no difference in performance and supporting an overfitting story. Conversely, when comparing Sharpe ratios of value in each asset class to that from the original asset class in which it was discovered (US equities), the other asset classes show similar or better performance, suggesting that the value factor was not overfitted to US stocks.

Repeating the analysis for momentum in Panel B, we also find that momentum's performance is weaker out-of-sample (about 23% worse), but the differences are insignificant as we fail to reject the F-test that the in- and out-of-sample performances are the same (p-value $= 0.09$). Relative to the original asset class for momentum (US stocks) the other asset classes produce about half of the momentum profits.

For the carry factor (Panel C), the out-of-sample performance is about three-quarters that of the original sample, and the F-test of equal performance fails to be rejected (p-value of 0.24). For currencies, the original asset class in which carry was discovered, the post-sample carry performance is larger than the original sample performance, and the average performance of carry outside of currencies is almost twice as large, suggesting that the carry factor is not overfitted to currencies.

Panel D reports results for the defensive factor. Here, the out-of-sample performance of the defensive factor is equivalent or larger than in the original sample for every asset class, which is inconsistent with overfitting bias. However, defensive strategies do perform better in individual equities, where they were first discovered.

Panel E shows the results for a diversified multifactor portfolio in each asset class. Defining the subperiod partitions here is tricky since different factors have different sample periods of discovery and to make meaningful comparisons, all four factors should be in the portfolio in every subperiod. Taking the union of original sample periods (1960 to 2009), or their intersection (1973 to 1981) has pluses and minuses, since the former is likely too long a sample and includes factors both before and after their discovery, whereas the latter is too short a sample and missing significant parts of the original sample for most factors. As a compromise, we define the "original" in-sample period to be from 1960 to 1990 for the multifactor portfolio, which covers the majority of the in-sample periods for all factors, noting this definition is imperfect, and hence that the interpretation of the results for the multifactor portfolio is less clear. Nevertheless, the results echo a summary of our findings, with the out-of-sample Sharpe ratio being 18% lower and statistically significant (p-value $=$ 0.00).

Overall, we find robust out-of-sample evidence of factor premia, rejecting that they are the result of spurious data mining, but find that overfitting biases may contribute to a significant decline in the out-of-sample efficacy of these factors.

2.2 *Variation due to arbitrage activity?*

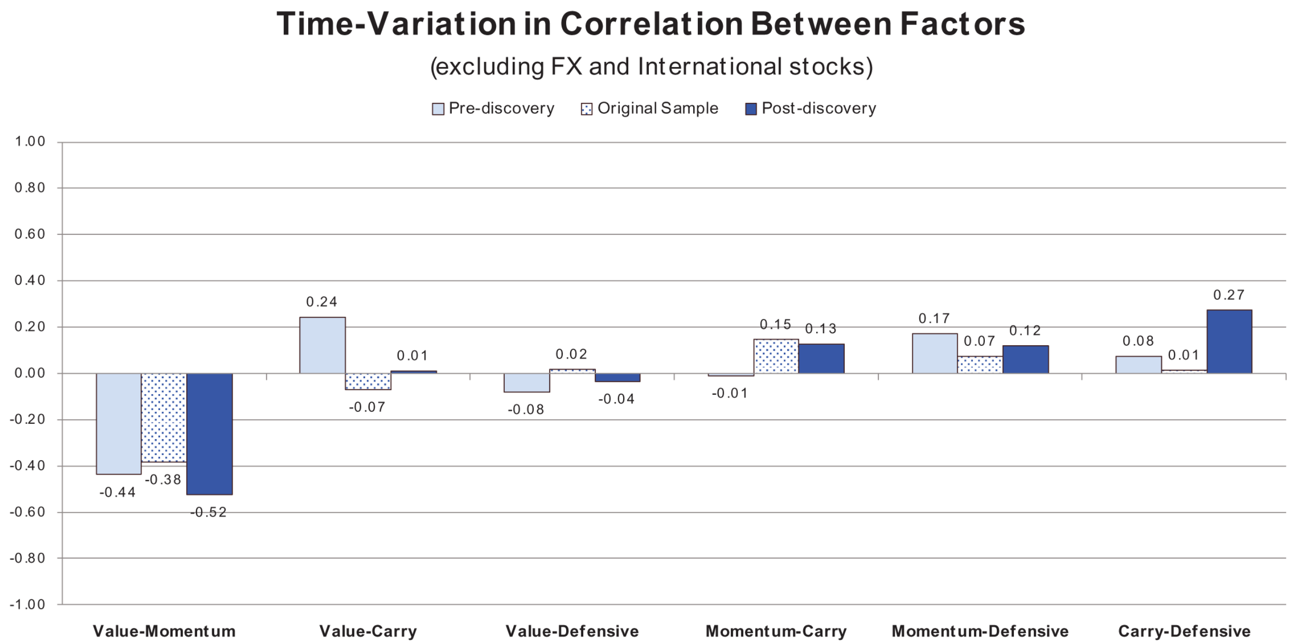
We turn next to an assessment of whether informed trading from arbitrageurs contributes to factor return variation over time. To control for the confounding effects of overfitting, we examine the difference in factor performance between the out-of-sample post- and pre-sample periods, where the latter should not be affected by arbitrage activity.

Panel A of Table 2 shows that value in US equities, equity indices, and fixed income does better in the pre-sample period than in the post-sample period, consistent with value degradation from informed trading, though the differences are insignificant. However, for value in commodities, the post-sample period produces *better* performance than the pre-sample period. And, in international equities and currencies (which do not have a pre-sample period) the post-sample performance of value is as large as the original sample. These results are inconsistent with arbitrage activity or informed trading reducing the value premium post-publication. A formal test of whether the post-sample performance is the same as the pre-sample performance fails to reject (p-value of 0.764).

Panel B similarly finds no evidence that momentum is weaker in the post-sample period relative to the pre-sample period, which challenges the notion that informed arbitrage activity drives down momentum's performance. The difference between the pre- and post-sample performance of momentum is no different from zero (p-value = 0.696). Panel C similarly shows no post- versus pre-sample performance difference for carry (p-value = 0.313), and the post-sample performance for defensive (Panel D) is actually greater than the pre-sample evidence, further challenging the arbitrage story. For the multifactor portfolios (Panel E), there is no difference in pre- versus post-sample performance (p-value = 0.834). The results do not support informed trading impacting the efficacy of these factors or contributing to time variation in factor returns.

Contrary to McLean and Pontiff (2016), we find little evidence that these factors were affected by informed trading after their publication. We acknowledge, however, that the power to detect arbitrage crowding into a factor and its effect on prices is challenging (Alquist *et al.*, 2019) and that our sample and tests are decidedly different from McLean and Pontiff (2016), who study 97 factors in US equities from 1963 to 2014 only. Nevertheless, we offer an alternative test of factor return degradation due to informed trading that has substantially more statistical power, yet does not show much support in the data. Our findings at least soften the conclusion that arbitrage activity has led to a reduction in factor return premia, given that we find no evidence for it across a variety of factors, asset classes, and time periods.

Another implication from increased arbitrage activity post-publication is excess correlations due to price pressure from trading (Lou and Polk, 2021; Alquist *et al.*, 2019). Changing correlations can also provide another test of data mining, as Linnainmaa and Roberts (2018) argue that data snooping biases can artificially lower correlations of factors in-sample. To test these implications, we examine the correlations in the pre-, original, and post-sample periods. Overfitting implies weaker out-of-sample correlations across asset classes for a given factor, while informed trading implies stronger correlations among factors and across asset classes post-discovery.



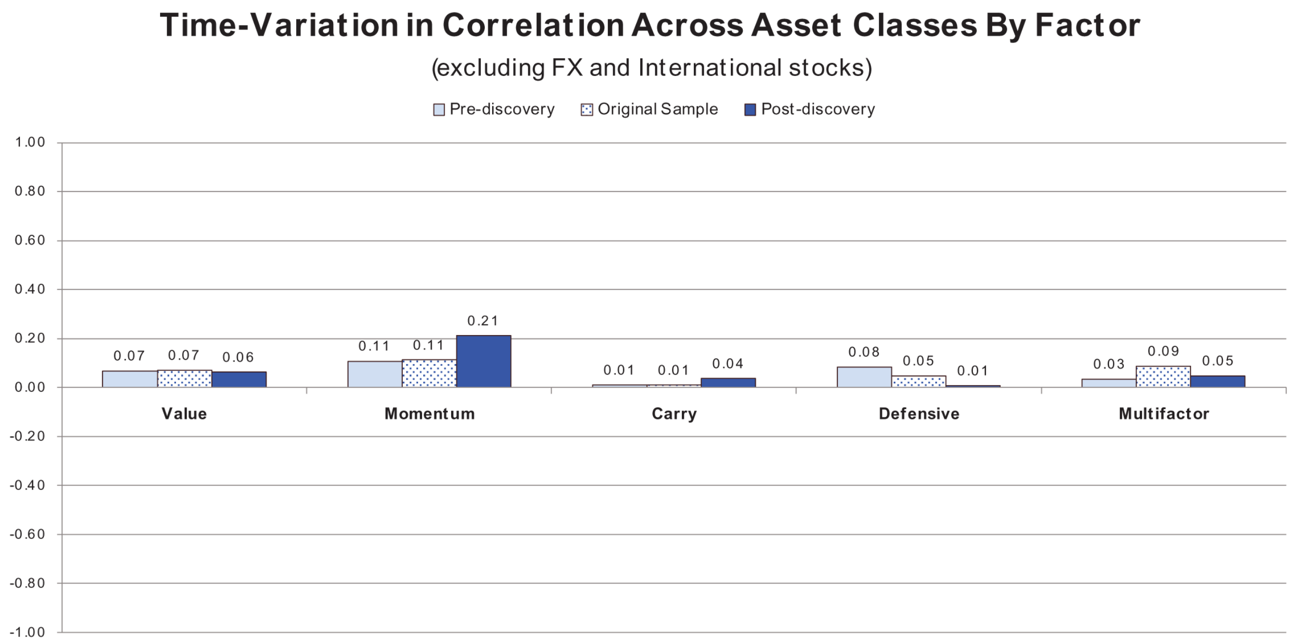


Figure 2 Correlations in the original-sample, post-publication, and pre-sample periods.

The first graph plots time variation in the correlation *between* factors (e.g., correlation between value and momentum). We examine all pairwise correlations between the four factors applied across all asset classes simultaneously, and estimate them separately over the pre-, original-, and post-sample periods, which are all reported below. The second graph examines correlations *across* asset classes for a given factor, by averaging the pairwise correlations between asset classes for a given factor (e.g., the average correlation of the value factor across markets). We exclude currencies and international stocks that have the shortest sample periods.

Figure 2 plots time variation in the correlation between and across factors. There is little variation in the correlations across the subperiods. In the first graph, correlations across factors are no higher in the out-of-sample periods than they are in the original sample. Moreover, there is no evidence that correlations across factors are higher post-factor discovery, as suggested by increased arbitrage activity. In the second plot, we find no evidence that correlations across asset classes for a given factor are higher post-discovery either. Hence, both the diversification benefits across factors and across asset classes do not appear to be different in the pre- versus post-sample periods. The evidence is inconsistent with arbitrage activity affecting the factors meaningfully.

3 Macroeconomic Exposure

Given the variation in factor return premia per-unit-of-risk we find in Section 1, and the ability of overfitting and informed trading to explain only a portion of it, we consider here what other sources of variation might be driving these dynamics. We examine a variety of economic shocks and news motivated by various asset pricing theories. Past attempts to link factors to economic risks have proven challenging due to limited time series. Our much longer and broader sample offers 50 years of additional economic events across different markets to help identify these relationships and test previously documented relationships out-of-sample. The breadth of asset classes also helps reduce noise that may cloud these relationships. This examination also provides a stronger test of dynamic asset pricing theories (Merton, 1973) using a wealth of novel data.

3.1 Factor return exposure

Table 3 reports results from a time series regression of each factor's returns over the last century on various economic measures. The first variable we examine is a measure of illiquidity risk from Amihud (2002) and Acharya and Pedersen (2005). Pastor and Stambaugh (2003), and Sadka (2006) show that individual stock momentum is exposed to liquidity risk and Asness *et al*. (2013) show some evidence that value and momentum across all asset classes are oppositely exposed to liquidity shocks. The second variable we examine is the Baker and Wurgler (2006) sentiment index, which Baker and Wurgler (2006) show explains variation in value, momentum, and other equity factor returns over time. We also include equity market volatility (realized volatility of the global equity portfolio over the prior 36 months) as a proxy for arbitrage costs and uncertainty in markets.

In an attempt to link factor returns to macroeconomic models (Breeden, 1979; Campbell and Cochrane, 1999; Lettau and Ludvigson, 2001, 2009; Bansal and Yaron, 2004; Lewellen *et al*., 2010; Greenwald *et al*., 2014), we look at measures of macroeconomic activity. An issue with looking at macro variables is that there is little theoretical guidance on how factors *should* interact with the macroeconomy. While macroeconomic variables have an intuitive link to long-only asset classes like equity and bond markets, their impact on long-short factors is not obvious. Some studies link value to long-run consumption growth (Parker and Julliard, 2005; Hansen *et al*., 2008; Malloy *et al*., 2009) and others tie value and defensive factors to discount rate sensitivity possibly related to interest rate regimes (Lettau and Wachter, 2007; Gormsen and Lazarus, 2019). Investment-based theories (Cochrane, 1991, 1996; Gomes *et al*., 2003; Carlson *et al*., 2004; Zhang, 2005; Xing, 2008; Li *et al*., 2009; Liu *et al*., 2009; Belo, 2010; Li and Zhang, 2010; Cooper and Priestley, 2011; Liu and Zhang, 2014; Hou *et al*., 2015) link asset pricing factors to economic shocks that impact firm investment, which can be related to the business cycle, interest rates, growth, and even political

Table 3 Contemporaneous and lagged macroeconomic exposures.

	Value	Momentum	Carry	Defensive	Multifactor
Panel A: Contemporaneous economic activity					
Amihud illiquidity risk	0.0031	-0.0031	0.0006	0.0036	0.0004
	(1.95)	(-1.48)	(0.44)	(2.16)	(0.54)
Baker-Wurgler sentiment	0.0014	-0.0008	0.0007	**0.0023**	0.0010
	(2.22)	(-0.99)	(1.20)	(**3.47**)	(3.17)
Equity market volatility	0.0137	-0.0131	0.0749	-0.0205	0.0044
	(0.32)	(-0.23)	(1.84)	(-0.45)	(0.20)
GDP growth	-0.0519	0.0624	0.0197	-0.0461	-0.0094
	(-1.25)	(1.15)	(0.51)	(-1.06)	(-0.45)
CPI inflation changes	0.0298	0.0621	0.0310	-0.0659	0.0310
	(1.00)	(1.58)	(1.11)	(-2.09)	(2.05)
Tail risk dummy	0.0003	0.0012	-0.0009	-0.0051	-0.0005
	(0.17)	(0.56)	(-0.57)	(-2.85)	(-0.58)
Geopolitical risk index	-0.0000	-0.0000	-0.0000	0.0000	-0.0000
	(-0.28)	(-0.61)	(-1.22)	(0.12)	(-0.99)
Real interest rate	0.0097	0.0309	0.0128	-0.0268	0.0073
	(0.27)	(0.66)	(0.38)	(-0.72)	(0.41)
One-year change in real rate	-0.0121	-0.0054	0.0484	0.0807	0.0289
	(-0.34)	(-0.12)	(1.46)	(2.16)	(1.61)
Slope of yield curve	0.0222	-0.0307	0.0850	0.0163	0.0276
	(0.28)	(-0.30)	(1.15)	(0.20)	(0.69)
One-year change in slope	0.0364	-0.0105	-0.0030	0.0074	0.0067
	(0.62)	(-0.14)	(-0.05)	(0.12)	(0.23)
Contraction dummy	0.0020	0.0017	-0.0031	0.0045	0.0011
	(0.49)	(0.32)	(-0.82)	(1.04)	(0.55)
Expansion dummy	0.0011	-0.0021	0.0009	-0.0033	-0.0007
	(1.01)	(-1.51)	(0.87)	(-2.96)	(-1.35)
Slowdown dummy	0.0062	-0.0034	0.0018	-0.0033	-0.0009
	(1.97)	(-0.82)	(0.62)	(-1.00)	(-0.58)
R^2	6.4%	2.8%	2.9%	9.4%	4.7%
Panel B: Lagged economic news					
Amihud illiquidity risk	0.0032	-0.0041	0.0003	0.0021	0.0001
	(1.98)	(-1.95)	(0.17)	(1.22)	(0.12)
Baker-Wurgler sentiment	0.0013	-0.0009	0.0006	0.0019	0.0009
	(1.98)	(-1.10)	(1.00)	(2.85)	(2.75)
Equity market volatility	-0.0112	0.0096	0.0862	0.0407	0.0184
	(-0.25)	(0.17)	(2.08)	(0.87)	(0.83)

Table 3 (*Continued*)

	Value	Momentum	Carry	Defensive	Multifactor
GDP growth	-0.0653	-0.0098	0.0231	-0.0047	-0.0300
	(-1.55)	(-0.18)	(0.59)	(-0.10)	(-1.42)
CPI inflation changes	0.0286	0.0916	0.0384	-0.0500	0.0398
	(0.95)	(2.34)	(1.37)	(-1.57)	(2.64)
Tail risk dummy	-0.0012	-0.0010	-0.0010	**-0.0074**	-0.0024
	(-0.68)	(-0.45)	(-0.61)	**(-4.05)**	(-2.82)
Geopolitical risk index	0.0000	-0.0000	-0.0000	0.0000	-0.0000
	(0.66)	(-2.24)	(-0.31)	(0.21)	(-1.04)
Real interest rate	0.0112	0.0856	0.0206	-0.0163	0.0251
	(0.32)	(1.86)	(0.63)	(-0.44)	(1.42)
One-year change in real rate	-0.0009	-0.0194	0.0340	0.0396	0.0214
	(-0.03)	(-0.41)	(1.01)	(1.03)	(1.18)
Slope of yield curve	0.0044	0.1003	0.1340	0.0378	0.0695
	(0.06)	(0.99)	(1.85)	(0.46)	(1.78)
One-year change in slope	-0.0058	-0.0119	-0.0048	0.0413	-0.0077
	(-0.10)	(-0.15)	(-0.09)	(0.66)	(-0.26)
Contraction dummy	-0.0010	-0.0085	-0.0085	0.0004	-0.0048
	(-0.24)	(-1.59)	(-2.23)	(0.08)	(-2.32)
Expansion dummy	0.0008	-0.0012	0.0012	-0.0013	-0.0001
	(0.73)	(-0.83)	(1.15)	(-1.12)	(-0.13)
Slowdown dummy	0.0069	-0.0125	0.0008	0.0024	-0.0024
	(2.16)	(-3.03)	(0.28)	(0.71)	(-1.53)
R^2	5.6%	4.0%	3.4%	8.1%	6.1%

The table reports results from time series regressions of each factor's returns over the last century on various economic measures. Panel A reports contemporaneous regressions between factor returns at time t and the economic variables at the same time t. The variables include the illiquidity risk variable of Amihud (2002), the Baker and Wurgler (2008) sentiment index, equity market volatility (realized volatility of the equal-weighted country indices, estimated over the prior 36 months), global GDP growth (growth over the last year averaged over the US, UK, Germany, and Japan), global CPI inflation growth (growth over the last year in inflation averaged over the US, UK, Germany, and Japan), a tail risk indicator (if the developed equity market index is in the lower fifth percentile), geopolitical risk index (from http://www.policyuncertainty.com/gpr.html), the current real interest rate (3-month short rate minus expected inflation), 1-year change in the real rate, the current slope of the yield curve (10-year bond yield minus 1-year bond yield), the 1-year change in the slope of the yield curve, and three business cycle indicators: contraction, expansion, and slowdown, which are determined using levels and changes in GDP growth. We define periods into positive and negative growths based on GDP growth each quarter, and also define periods into "accelerating" and "decelerating" growth each quarter based on the change in GDP growth. The intersection of these two indicators creates four subperiods: contraction (negative growth and negative change in growth), recovery (negative growth and positive change in growth), expansion (positive growth and positive change in growth), and slowdown (positive growth and negative change in growth). Panel B lags all macroeconomic variables by one period to capture their announcement lag, which captures the news associated with the economic variable and its contemporaneous impact on markets. **Bold** highlights coefficients that pass statistical significance after the Bonferroni-adjustment for multiple comparisons.

uncertainty. For example, Berk *et al*. (1999) and Johnson (2002) tie momentum to growth options of the firm. Hou *et al*. (2015) tie value, momentum, and quality factors to investment. Some theories, however, make sharper predictions. For instance, the duration-based asset pricing theories of Lettau and Wachter (2007) and Gormsen and Lazarus (2019) argue that the value factor, which tends to be short cash flow duration, is sensitive to discount rate shocks and hence more exposed to a loosening of monetary policy, all else equal. Rare disaster theories (Tsai and Wachter, 2015; Gabaix, 2012) and downside risk (Lettau *et al*., 2014) posit that factor returns decline with tail events and expected returns rise with the probability of tail events.

The existence of the same factor premia in other asset classes besides equities challenges many of these models, which are equity-centric theories. Given the abundance of theories making links between factor returns and macro variables, and the lack of theory tying macro variables to factor premia in other asset classes, we embark on an empirical exploration of factor exposures across asset classes to a host of macroeconomic variables.

We use global GDP growth (real growth over the last year averaged over the US, UK, Germany, and Japan) and global CPI inflation rate (inflation rate over the last year averaged over the US, UK, Germany, and Japan), which are two variables that Chen *et al*. (1986) find matter for stock returns. We proxy for tail events using a binary indicator that equals one if the developed equity market return is in the bottom 5th percentile of its historical return distribution and zero otherwise. We use a geopolitical uncertainty risk index from http://www.policyuncertainty.com/gpr.html, following the methodology of Baker *et al*. (2016) and Caldara and Iocoviello (2018). To capture interest rate environments, we look at the level and one-year changes in the real interest rate (3-month short rate minus expected inflation, where the latter is a three-year moving average of inflation) and the level and one-year changes in the slope of the yield curve (10-year minus 3-month risk-free interest rate). To capture business cycle variation, we divide periods into positive and negative growth based on annual GDP growth each quarter, and define periods into "accelerating" and "decelerating" growth based on the change each quarter in the year-on-year GDP growth number. The intersection of these indicators creates four subperiods: contraction (negative growth and negative change in growth), recovery (negative growth and positive change in growth), expansion (positive growth and positive change in growth), and slowdown (positive growth and negative change in growth). Regime change is triggered by GDP growth or changes in GDP growth moving at least $+/-1$ standard deviations based on a 10-year rolling historical window, to avoid frequent switches between regimes. Unlike NBER recession and expansion dates, these subperiods are determined solely based on ex ante information. Appendix A details the construction of these variables and their data sources.

An important aspect of some of the macroeconomic variables is their timing. Many macroeconomic variables are announced and reported after the actual quarter or month they pertain to. For example, the initial estimate of the second quarter GDP only gets announced in July (third quarter) of the same year. A question arises then as to whether we should match second quarter financial returns to second quarter GDP numbers, which reflected actual GDP growth at that time, or whether we should match third quarter financial returns to the second quarter GDP number since the market learns about the second quarter growth only in the third quarter. The first choice measures the relationship between returns and *actual* economic activity. The second choice measures the relationship between returns and *news* of economic activity, which the market finds out later. We examine both. Panel A of Table 3 examines the contemporaneous relationship between the factor returns and economic activity, and Panel B lags all variables by one period to capture news.

The first column of Panel A of Table 3 reports the results for the value factor across all asset classes. Value loads positively and significantly on illiquidity risk, the Baker and Wurgler (2006) sentiment index, and the economic slowdown indicator, which are broadly consistent with Asness *et al*. (2013) and Baker and Wurgler (2006), respectively, over shorter sample periods. However, none of these coefficients are statistically significant after accounting for multiple comparisons. For momentum, nothing is significant. Similarly, carry returns have no contemporaneous relation to any of the variables. Defensive, like value, has higher returns when illiquidity risk and sentiment are high. In addition, one-year changes in the real rate affect defensive strategies positively and tail risk, inflation, and business cycle expansions affect defensive strategies negatively, which is intuitive. However, only sentiment is significant after accounting for multiple testing. We find little evidence that factors vary with interest rate environments in a manner purported by duration models (Lettau and Wachter, 2007; Gormsen and Lazarus, 2019).

Panel B of Table 3 examines the regressions using the same independent variables lagged an additional period (which can be a month or a quarter, depending on the frequency of the variable as detailed in Appendix A). For some variables, like illiquidity risk, volatility, and sentiment, lagging the variables represents "news" in the sense that it is the most recent information an investor could obtain in real time about these variables. For many of the macroeconomic variables, lagging ensures that the actual news was released by the portfolio formation date. The coefficients on these variables represent predictive relationships. The evidence for economic news predicting factor returns is as weak as contemporaneous activity, with low R-squares and insignificant coefficients.

Overall, there is little evidence that the factor returns vary in a meaningful way with macroeconomic variables, either contemporaneously or predictively. Despite our long and broad sample providing a rich set of macroeconomic events and added statistical power, we do not find much macroeconomic exposure for long-short factors. The results are broadly consistent with Griffin *et al*. (2003), Asness *et al*. (2013), and Herskovic *et al*. (2019), and are inconsistent with other studies (Chordia and Shivakumar, 2005; Hodges *et al*., 2017) that examine much shorter histories and equity-only factors. Factor investing does not appear to be afflicted by the same macroeconomic risks that move general stock and bond markets, and are therefore diversifying to traditional asset allocation strategies.

3.2 *Factor correlations*

Although macroeconomic shocks do not appear related to factor returns, they may be related to factor risks and correlations. Figure 3 plots the correlations between factors (first graph), as well as the correlations across asset classes for a given factor (second graph), in various economic environments. The first graph plots the correlations between factors in the 20% worst and best equity, bond, and global market environments, the 20% highest and lowest market volatility periods, as well as during recessions and expansions. We see mixed behaviors across style pairs. The correlation between value and momentum is −0.66 during the best performing market months and −0.41 during the worst performing market months. This result implies

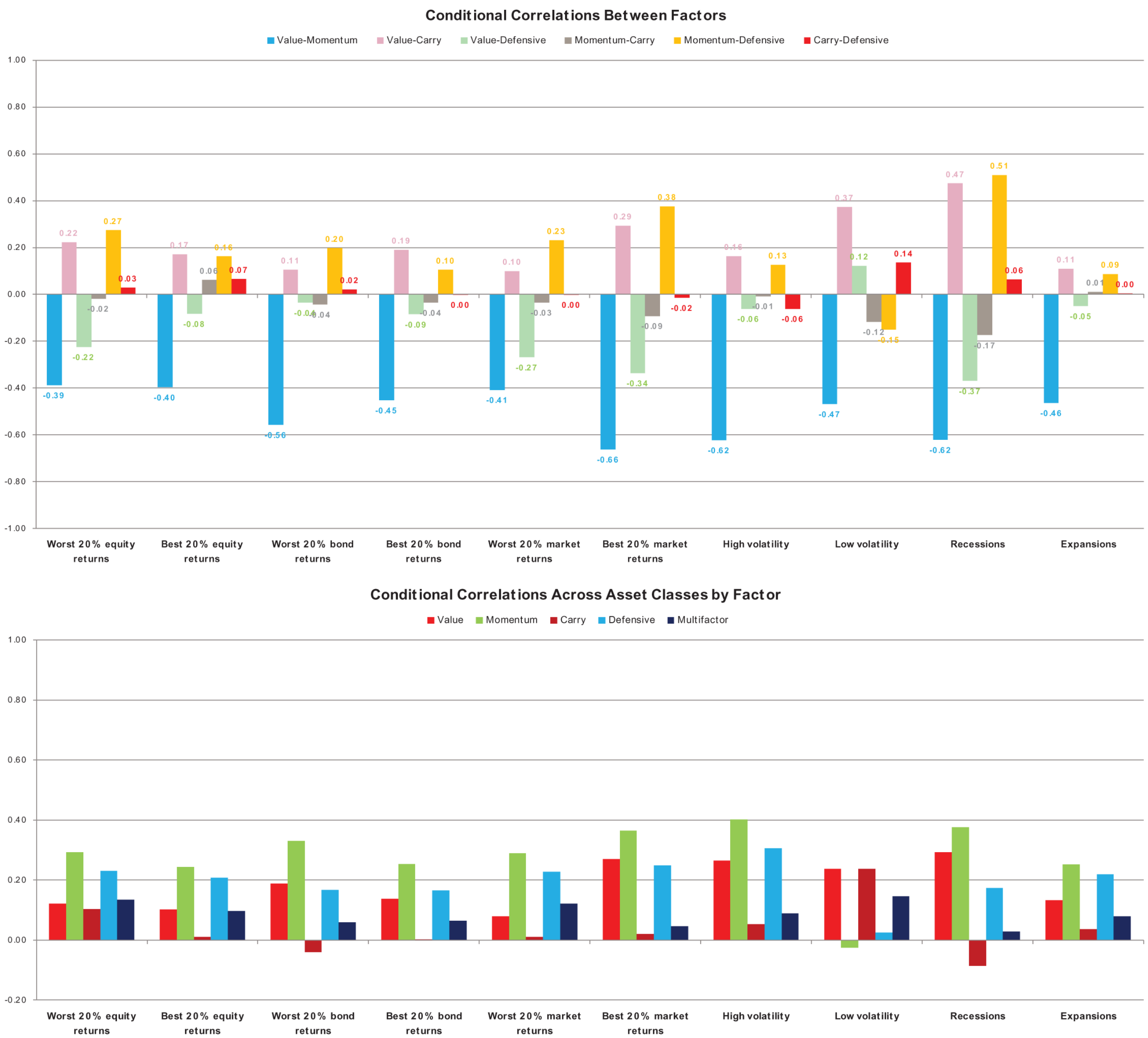


Figure 3 Correlations in different economic environments.

The figure plots the correlations between factors and the correlations across asset classes for a given factor in different economic environments. We separately compute correlations for the 20% worst and best months of global equity returns (using the MSCI index), the 20% worst and best months of global bond returns (using the Barclays Aggregate Bond Index), the 20% worst and best global market returns (using a volatility-weighted average of all asset classes that includes stocks, bonds, stock indices, currencies, and commodities), the top and bottom 20% of months based on equity market volatility (realized volatility over the last 36 months), as well as during global recessions and expansions using the NBER's business cycle definitions applied to all developed markets in our sample. The first graph plots the six pairwise correlations between factors across all asset classes in each economic environment. The second graph repeats the same exercise looking at average pairwise correlations across asset classes for a given factor.

that value and momentum are even better diversifiers for each other during good times and less so during bad times.[16] In contrast, value and carry and momentum and defensive exhibit higher correlation during good times than bad times. The correlations between value and defensive, momentum and carry and carry and defensive are roughly the same in up and down markets.

There are also interesting correlation dynamics across high- and low-volatility environments. With the exception of momentum and carry and momentum and defensive, all other pairs are more diversifying to each other in high-volatility periods. Finally, some interesting patterns emerge when comparing recessions versus expansions. Value and momentum, value and defensive, and momentum and carry all provide bigger diversification benefits during recessions, while value and carry and momentum and defensive are more correlated with each other during recessions.

The second graph in Figure 3 plots the average correlation across asset classes for a given factor. For value and momentum, the correlations across asset classes are lower during the worst market return months, indicating that diversification benefits across asset classes are better during bad times. However, during high volatility periods, momentum and defensive have much stronger cross-asset correlations than during low-volatility times, whereas for carry it is the opposite. During recessions, there is more cross-asset correlation for value and momentum factors, but slightly less cross-asset correlation for carry and defensive factors.

Taking all of these dynamics in correlation structure together, the multifactor portfolio exhibits slightly higher cross-asset correlation in down markets, low-volatility environments, and expansions, but the differences are small. Overall, we do not find much impact on returns from macroeconomic shocks, but do find some variation in correlations and risk associated with macroeconomic regimes.

4 Conditional Factor Premia and Factor Timing

Given the variation in factor returns and risks, can we find conditioning information that can predict risk-adjusted returns to the factors?

Predictive tests using conditioning information have been the workhorse for assessing dynamic asset pricing models (Hansen and Richard, 1987). We seek to capture predictable time-varying risk-adjusted factor premia through factor-timing strategies. More broadly, this study sheds light on the vast literature of factor timing and conditional returns. Since the same unconditional factor premia exist across asset classes, it is interesting to assess whether similar *conditional* premia exist across asset classes. Our long and broad sample provides a more powerful laboratory to detect conditional premia and to test the robustness of previous timing studies.

Timing studies typically fall into one of three categories: (1) a single factor in a single asset class timed with a single predictor,[17] (2) multiple factors in a single asset class timed with one or many predictors,[18] and (3) a single factor in multiple asset classes timed with a related predictor.[19] We expand the evidence on timing and return predictability by examining multiple factors across multiple asset classes using multiple timing signals (and methods) to synthesize factor timing across markets. To assess the magnitude of conditional factor premia and make comparisons across predictors and methods, we compare timing models on the basis of returns with an implementable trading strategy. This approach has several benefits. First, it puts all factor-timing models on equal footing by comparing them based on out-of-sample return performance. Second, returns per dollar of exposure provide a measure of economic magnitude. Third, returns allow us to assess the marginal benefit of factor timing to an investor's optimal static factor portfolio. Fourth, focusing on the returns to an investment strategy circumvents problems with other timing metrics such as *R*-squares, where degrees of freedom, precision of the parameters, and other statistical biases can cloud inferences (Stambaugh, 1999; Boudoukh *et al*., 2020). The returns to an investment strategy encapsulate these issues because poor model estimation will generate poor out-of-sample performance. Finally, unlike other metrics such as *R*-squares, we do not need to specify a benchmark for what constitutes meaningful predictability, since the natural benchmark to any risk-adjusted trading strategy is zero.

4.1 Timing signals

We begin with the best-known factor-timing signal: valuation spreads. Valuation ratios have long been used to forecast equity market returns, dating back to Fama and French (1988) and Campbell and Shiller (1988). A natural candidate for this metric is a value measure like aggregate book-to-price or CAPE to indicate when expected returns are high or low, which can be driven by time-varying risk premia or mispricing from investor sentiment. The same concept can be applied to long-short factors by comparing the average valuations of long positions with that of short positions. The difference, often referred to in the literature as the value spread,[20] may be informative about the conditional expected return of the factor.[21] If valuations of the factors are indicative of time-varying premia, then we expect a positive relationship between the factor's valuation and future returns.

While valuation spreads are an intuitive and prominent timing predictor, other variables have been used to provide information about conditional expected returns. Gupta and Kelly (2019), Ehsani and Linnainmaa (2019), and Arnott *et al*. (2019) use factor momentum, or the past return on the factor itself, to time US equity factors. Another conditioning variable, similar in spirit to valuation or momentum timing, is to use the spread in the factor characteristic itself to time factors. The idea is that when the spread in characteristics between the long and short positions is particularly wide, this may be when the factor's return premium is expected to be larger. For the value factor, the "characteristic spread" is the same as the valuation spread. For other factors, the characteristic spread represents the carry of the carry factor (as in Koijen *et al*., 2018), the momentum of the momentum factor, and the beta spread of the defensive factor.

We also look at five-year reversals as a timing signal, which is the negative of the past five-year return on the factor. In addition, we examine volatility timing following Moreira and Muir (2017) using the inverse of the standard deviation and variance of each factor as conditional information. We also examine a host of business cycle and macroeconomic variables to time factors, including GDP growth and inflation growth, as well as the business cycle indicators. Finally, we add two general market timing variables: CAPE and the lagged realized volatility of the market, "VIX."

4.2 Timing methodologies

There are many ways to use timing signals to construct portfolios. We focus on four methodologies (though have looked at 19 variants in unreported results). The first is a simple z-score methodology that standardizes the signal to a mean zero, unit variance measure over its history (standardization also allows us to compare and combine different timing signals). We time the factor by increasing its weight at each point in time in proportion to its z-score, with the sign and magnitude of the z-score determining the dollars to be long or short. Since many of the signals experience extreme values, we use the historical median and absolute deviation from the median to define "z-scores" and cap the weights at $+2$, -2. Also, to ensure no look-ahead bias, the z-scores are estimated using an expanding window from the first return observation in the sample to month $t-1$, requiring at least 10 years of history, and then applied to factor returns at time t.

The second methodology is a predictive regression of factor returns on lagged timing signals to measure their empirical relation and use the coefficients to time factors. Next, we also impose economic restrictions on the coefficients as suggested by Campbell and Thompson (2007). Finally, some studies use in-sample moments to determine parameters, which is not fully implementable in real time and may generate a look-ahead bias. We look at this method, too, to estimate the bias in using full-sample parameters.

This analysis serves as a grand specification search for the best way to extract predictive content from the signals. In total, we analyze 11 timing signals across 19 different specifications, for all 20 factor-by-asset class long-short portfolios plus the six multifactor and multi-asset portfolios, creating $11 \times 19 \times 26 = 5{,}434$ timing strategies. While this search raises the specter of data mining, given the lackluster findings we will show, the aim is to evaluate the robustness of those weak results. If a massive search for timing turns up little, then we feel more confident that the lack of strong results is not simply unlucky, but rather, evidence that factor timing is challenging. Given the number of timing methods and signals we tried, any significant results must be balanced against the number of comparisons being made, in this case a Bonferroni correction to the normal 5% significance level results in a 0.0009% threshold.

Figure 4 reports the results for all 11 factor-timing variables across the four core timing model specifications for the multifactor, across-all-asset-classes portfolio. Looking at factor timing across all asset classes and all factors simultaneously provides the most flexibility and power to detect conditional premia in the data. Since timing strategies can increase the unconditional betas to the static factors (Asness *et al.*, 2017, 2018), we regress the returns of the timing strategy on the static factors, where the alpha represents pure timing returns stripped of any unconditional exposure to the factors.

The information ratio (alpha per unit of residual risk) of value timing of the factors across asset classes ranges from 0.25 to 0.50. This result is directionally consistent with the results in Cohen *et al*. (2003) and Asness *et al*. (2000), who find positive results to valuation timing of the US equity value factor over their shorter sample periods using slightly different methodologies. Baba *et al*. (2019) find that returns to value strategies are predictable by the value spread, though they focus predominantly on equity-only factors. Asness *et al*. (2021) show that "deep value" periods, where the valuation spread between cheap and expensive securities is in the extreme 20th percentile relative to its history, predict value factor returns in global stocks, equity index futures, currencies, and bonds. Brooks and Moskowitz (2018) show that value spreads predict the returns of global bond portfolios. Our findings over a much longer period provide out-of-sample evidence on valuation timing.

The next four bars in Figure 4 report the results for factor momentum (past 12-month return on the factor), which delivers weaker timing performance. The results are directionally consistent, but weaker than Gupta and Kelly, 2019; Arnott *et al*., 2019. The main differences between our study and theirs are (1) they focus on several

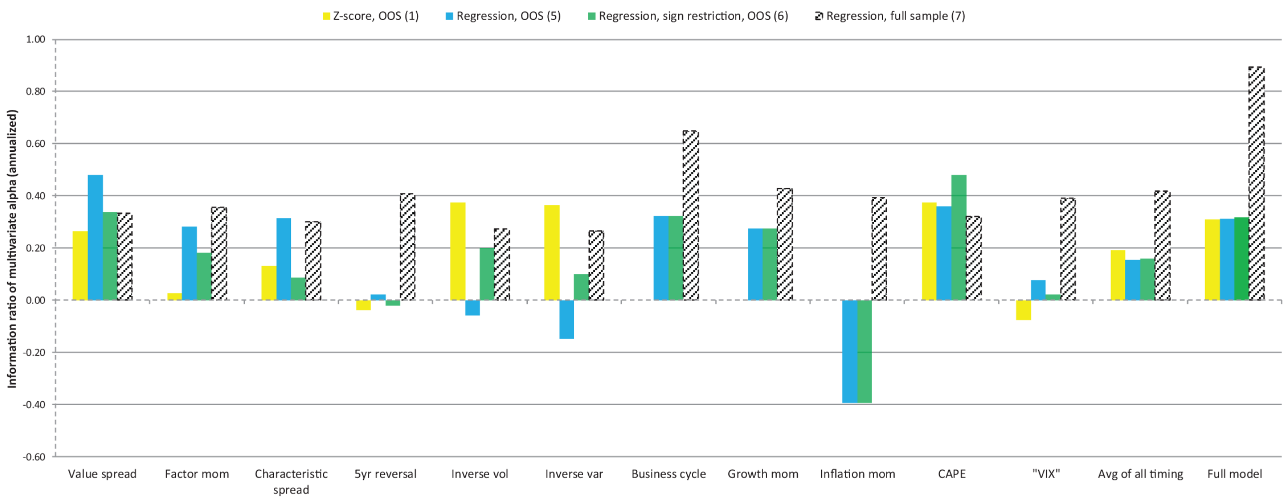


Figure 4 The returns of factor timing.

The figure reports timing portfolio information ratios relative to the multivariate static factors in each asset class for different timing variables using different timing methodologies. We use the following timing methodologies: z-score, out-of-sample regression with no sign restrictions, out-of-sample regression with an economic sign restriction on the timing variable (e.g., value spreads should have a positive coefficient), and a full-sample regression (in-sample) that places no restrictions on any coefficients. The timing variables are the value spread, past 12-month return on each factor (factor momentum), average characteristic of the factor itself (characteristic spread), which is the value spread for value portfolios, the momentum for momentum portfolios, the average carry for the carry portfolios, and the average negative beta for the defensive portfolios, 5-year reversals (negative of past 5-year return on the factor), inverse volatility (where volatility is estimated over the prior 36 months of returns), inverse variance, business cycle (an ex ante measure that seeks to identify stages of the business cycle—contraction, recovery, slowdown, expansion—where we use both the level and change of GDP growth, compute a rolling 10-year z-score of level and changes in GDP growth, and identify the turning point of a business cycle as whether the z-score breaks $+/-1.0$ to identify each of the four periods), growth momentum (moving average of annual GDP growth), inflation momentum (moving average of inflation growth), CAPE, and VIX (realized volatility of the market over the last 36 months). We also report a simple average of all the timing strategies as well as a timing strategy based on the "full model" that incorporates all timing variables into one model, under each methodology, to time the factors.

dozen long-short factors in individual US equities only, while we examine only four factors but study them across six different asset classes; and (2) we examine factor momentum timing over a much longer period.[22]

The next set of results in Figure 4 show that characteristic spread timing is slightly weaker than either valuation spread or momentum timing, suggesting that carry and defensive spreads are not as useful for timing. Five-year return reversals also do not produce any timing profits.

Moreira and Muir (2017) find that inverse volatility and variance predict conditional Sharpe ratios to US equity factors, as well as the currency carry factor. We examine inverse volatility timing for our factors across the six asset classes over the century, using the realized volatility and variance of each factor's returns over the prior 36 months. Figure 4 shows that the simple z-score timing model, which is closest to the method Moreira and Muir (2017) use, produces sizeable timing alphas, with information ratios of 0.37 and 0.36. The regression models do not fare as well, although placing economic restrictions on the coefficients helps significantly. The regression generates *negative* timing alphas out-of-sample when the coefficients are unconstrained. This evidence points to the importance of imposing theoretical restrictions on timing models.

The next set of results focus on business cycle and macroeconomic timing predictors. We use the business cycle variables described earlier to time the factors. One issue here is that theory provides little guidance on how these indicators should be related to future factor returns. This makes timing based on macroeconomic variables particularly challenging because you have to make two predictions—first forecast the macroeconomic event and second forecast what the factor exposures are to those macroeconomic events, which is not clear for market-neutral factors.

Given the lack of intuition for how factors should be affected by the macroeconomic variables, we cannot run the simple z-score timing methodology. We focus instead on the regression methodologies, although only those with no economic sign restrictions for the same reason. Figure 4 shows that there is some timing alpha from business cycle and growth momentum indicators, particularly using the full-sample regression to estimate the coefficients (unsurprising, given the look-ahead bias). Moreover, we showed in Table 3 that the in-sample estimated coefficients on the business cycle variables are not significant, and change sign across asset classes for a given factor. Timing based on growth and inflation momentum similarly show much weaker performance when parameters are estimated out of sample. The figure highlights the dangers of using in-sample parameter estimates, especially for theoretically ambiguous variables such as the macroeconomic measures.

The last two timing signals, CAPE and VIX, are designed to capture changing risk, risk aversion, or sentiment in the equity market. As Figure 4 shows, there is significant predictability for factor premia from CAPE. VIX, on the other hand, does not deliver any timing ability for the factors. The out-of-sample timing alphas are not reliably different from zero.

4.3 Full model timing

If each timing signal has some unique predictability for expected returns, and if there is independent information and independent error in the signals, then it may be more powerful to combine timing signals to better capture conditional return premia.

We report the equal-weighted average of all timing strategies, which shows small, but positive alphas with respect to the underlying static factors. However, a better way to combine the information across timing signals may be to look at them simultaneously to account for their interactions and marginal contributions. For the z-score methodology, we rank all assets based on their equal-weighted average z-score across all timing signals (excluding the business cycle variables that have no predicted sign). For the regression methodologies, we run a multivariate regression of future factor returns on all 11 timing signals and use the product of the estimated coefficients and the current variable realizations to produce an expected return forecast for each factor in each asset class. Under certain specifications, we also restrict the signs of the timing coefficients to match economic theory. The results for the full model show the most consistent timing returns. The out-of-sample performance delivers information ratios of 0.31 that are orthogonal to the underlying static factors. Imposing economic constraints on the coefficients slightly improves out-of-sample performance, generating an information ratio of 0.32. Using the full in-sample regression coefficient estimates for the timing model generates an information ratio of 0.89, which again highlights the dangers of using full-sample information. The results identify the presence of significant conditional factor return premia, which are better identified when combining a variety of timing signals.[23]

4.4 Economic impact of factor timing

What are the practical implications of our findings? We investigate here whether an optimal factor-timing portfolio is economically meaningful. We consider how much factor timing an investor would add to a static diversified factor portfolio in order to maximize the Sharpe ratio over the sample period. Table 4 reports the Sharpe ratios and information ratios (with respect to the underlying static factors), of various timing strategies. We compute the optimal ex-post timing weight on the factor-timing strategy that, when combined with the static multifactor, multi-asset portfolio, maximizes the unconditional Sharpe ratio. The first row reports statistics for the static diversified multifactor, multi-asset class portfolio that uses no timing, which has a Sharpe ratio of 1.48. We also report the annual two-sided turnover of this portfolio per dollar of leverage, which is 4.4 (i.e., long 220% and short 220% of net asset value). Turnover per dollar levered measures the amount of trading taking into account leverage to compare portfolios on the same scale.

The remaining rows of Table 4 report the same statistics for various timing strategies. We start with the timing strategy that yields the greatest profits—the full timing model using a full-sample regression whose parameters are estimated in-sample and impose no restrictions on the coefficients. As shown earlier, this strategy generates an information ratio with respect to the static underlying factors of 0.89. Combining this timing strategy with the static diversified factor portfolio produces a Sharpe ratio of 1.73, where the ex-post optimal weight on the timing strategy is 37.7%. These results represent the best-case scenario from timing in our sample using the timing signals we study and knowledge of the full-sample estimates of the parameters. Of course, this timing strategy is not implementable in real time because it requires knowledge of the full-sample parameters. Rather, this specification serves as a benchmark for an upper bound to the additional Sharpe ratio we can hope to gain from the conditional information in factors.

Timing strategies also require additional turnover and trading. Table 4 reports that this timing strategy has turnover per dollar leverage of 5.9, increasing the long-short turnover by 134%. Is the additional turnover worth it? Rather than attempt to build a trading cost model that applies to all asset classes we study, we instead back out the break-even costs from the additional turnover that would wipe out all of the gains from the timing strategy. Specifically, after determining the optimal weight on the timing strategy, we can compare two strategies: the “original” portfolio with constant strategic factor weights, and a second “timed” portfolio, which is an optimally weighted combination of the strategic and timing strategies. We then compute the difference in performance between the original and timed portfolios and compare this with the difference in their respective turnover levels to arrive at a calculation of how large the cost would have to be per dollar traded to offset the performance increase from adding timing.

In this particular case, that number is 7.3 basis points (bps) per dollar traded. As long as transactions costs are less than this, adding this timing strategy would improve the net returns of a diversified factor portfolio. Based on evidence in equities from Frazzini *et al*. (2020), trading costs at a reasonable size (e.g., 1% of daily volume) would be a little higher for a cost-efficient arbitrageur over the past two decades. Outside of equities, no good estimates of trading costs exist in the literature, but average trading costs in fixed income, currencies, and equity index futures might be a bit lower and commodities about the same or a bit higher, than in equities. We caution

Table 4 The economic impact of factor timing.

	In/Out-of-sample	Sharpe ratio	Information ratio	Turnover per $ leverage	Sharpe ratio static + timing (ex-post optimal)	Ex-post optimal weight on timing	Break-even trading cost per dollar traded for adding timing (in bps)
Static (no timing)	OOS	1.48	0.00	4.4	1.48	0.0%	
Timing strategies							
Full model, no restrictions	FS	0.87	**0.89**	5.9	1.73	37.7%	7.26
Full model, no restrictions	OOS	0.24	**0.31**	7.9	1.51	17.4%	1.79
Full model, economic sign restrictions	OOS	0.36	**0.32**	7.0	1.51	17.6%	1.86
Value spread	OOS	0.39	**0.34**	5.3	1.52	18.5%	2.07
Factor momentum	OOS	0.48	0.18	5.6	1.49	11.0%	1.04
Characteristic spread	OOS	0.16	0.09	5.0	1.48	5.5%	0.46
Five year reversal	OOS	−0.14	−0.02	5.1	1.48	0.0%	0.00
Inverse volatility	OOS	0.42	0.20	6.3	1.49	11.9%	1.10
Inverse variance	OOS	0.40	0.10	6.3	1.48	6.3%	0.52
Business cycle	OOS	0.21	**0.32**	5.5	1.51	17.9%	1.90
Growth momentum	OOS	0.09	**0.27**	5.9	1.50	15.7%	1.58
Inflation momentum	OOS	−0.22	−0.39	7.0	1.48	0.0%	0.00
CAPE	OOS	0.70	**0.48**	4.5	1.55	24.5%	2.99
“VIX”	OOS	−0.31	0.02	6.0	1.48	1.4%	0.10

Notes: The table reports the economic impact of factor timing using various factor timing signals. For each timing model, we report the Sharpe ratio, information ratio of the timing strategy relative to the underlying static factors, with bold numbers indicating significance at the 5% level, and the turnover per dollar leverage. We also report the ex-post optimal Sharpe ratio of combining the timing strategy with the underlying static factor portfolios, the ex-post weight placed on the timing strategy in that optimization, and break-even trading cost per dollar traded for adding the timing strategy to the static factor portfolios at the ex-post optimal weight. These statistics are reported for timing strategies that use the full model with no restrictions estimated over the full sample (FS), for the same model estimated out-of-sample (OOS) using an expanding window of data up to time $t-1$, and for the full model out-of-sample with economic sign restrictions placed on the coefficients. The last timing strategy is further broken down by each individual timing variable separately and the statistics reported for each timing variable. The static (no timing) factor portfolio across all asset classes is also reported for comparison in the first row.

that these estimates only reflect the current market and trading infrastructure; given our long study, investors would likely have faced much higher costs for the majority of our sample, and to reiterate, these are in-sample results with look-ahead bias.

The next row reports the same statistics for the same timing model that estimates all of its parameters out-of-sample. The performance of this out-of-sample timing model is weaker, generating an information ratio of 0.31. This timing strategy only improves the unconditional gross Sharpe ratio of the static diversified factor portfolio from 1.48 to 1.51, where the optimal weight on timing is 17.4%. The implied break-even trading cost from that increase in turnover is 1.8 basis points per dollar traded. Actual trading costs at a reasonable size likely exceed this figure. Imposing economic restrictions on the timing variables does little to improve the results, where the optimal weight on the timing strategy is 17.6% and the break-even trading cost of adding timing is 1.9 bps per dollar traded. The remaining rows of Table 4 report results for each timing signal separately. Value spreads, business cycle, growth momentum, and CAPE timing variables are the only ones that seem to improve out-of-sample performance. However, the added turnover from factor timing can only be justified if trading costs are minimal; on the order of 2–3 bps per dollar traded.

The case for adding factor timing to an already diversified multifactor portfolio is tenuous in practice. Despite looking at a plethora of timing strategies, methodologies, and signals, we find modest evidence of out-of-sample factor timing. Accounting for increased turnover and trading costs associated with factor timing, the net of cost returns to timing are likely de minimis.

On a more positive note, despite limited ability to profit from factor timing in real time, we find significant conditional return premia associated with common factors across diverse asset classes. A model that employs a slate of conditioning information performs better at forecasting conditional returns. This evidence offers hope for identifying conditional expected returns in the economy and the types of asset pricing models that can accommodate them. Future research may well uncover a more powerful way to extract conditional information that yields more substantial economic returns.

5 Conclusion

A century of data across six diverse asset classes provides a rich laboratory to investigate how canonical asset pricing factor premia vary over time. We find that return premia for value, momentum, carry, and defensive are robust and significant in almost every asset class over the last century, and vary significantly over time. Part of this variation comes from poorer out-of-sample performance from the original studies, consistent with overfitting biases. We find little evidence that these return premia have been altered by informed arbitrage activity. Examining a slew of macro variables and a variety of conditioning information and timing methodologies, we identify conditional risk-adjusted return premia to our factors over the century, finding the most consistent results when we impose theoretical restrictions and combine multiple sources of conditioning information. The evidence identifies significant, though modest, dynamic premia that are difficult to profit from. An optimal timing portfolio implementable in real time produces small profits once real-world implementation costs are considered. Hence, an investor should be cautious about deviating from a long-term static allocation to these factors through tactical factor timing.

Appendix A: Data Description and Sources

Global equity indices

We obtain returns on equity indices from 43 equity markets internationally from Global Financial Data, which include all countries covered in the MSCI World Index from 1920 to 2020. Since most countries have multiple equity indices at each point in time, we select the index that is investible, has the most representative coverage of the total stock market of that country (by market cap), and has the longest history. We use monthly index total returns (including returns from dividends) from Global Financial Data and subtract the local currency cash returns to get excess returns. We also use monthly futures returns from Bloomberg and Datastream, which covers a shorter history, to supplement these data.

Nine of the stock indices have data going back to the 1920s, most of the rest of the developed equity markets have index data going back prior to the 1950s, 1960s, or early 1970s, and emerging markets go back in some cases to the 1970s, most in the mid- to late-1980s, and two countries in 1991. All countries have returns up through November 2020, so the minimum history is 29 years (Poland) and the maximum is 100 years (AU, BD, FN, FR, SD, US, UK) of monthly returns.

Panel A of Table B1 reports summary statistics on the country indices covered, including the start dates for each country's index returns, the annualized mean and standard deviation of returns, and the worst 12-month return for each index over the sample period. The latter highlights some of the extreme events these markets have experienced (e.g., Germany, Brazil) over the last century that provides more tail events and downside risks to examine.

Global fixed income

We use nominal yield and total returns data of 10-year local currency government bonds as well as 3-month interest rates from Global Financial Data and supplement it with Bloomberg and Datastream. The cross-section of government bond indices includes 26 countries, covering North America, Western and Northern Europe, Japan, and the Antipodeans.

Given the evolving nature of bond issuance historically, the inputs into our bond yield and returns can vary over time. In general, 10-year bond yield and returns only become available between 1960 and 1980. Between 1920 and 1960, the database uses the closest available tenor to 10-year, while before 1920, the yield and returns are typically for individual bonds. Panel B of Table B1 reports summary statistics on the bonds. Again, there is both rich heterogeneity in bond returns across countries and significant extreme events that occur over our sample period.

Global currencies

We use spot and 1-, 2-, 3-, and 6-month forward exchange rates obtained from Bloomberg, Citi and interpolate the forward exchange rate for the next quarterly International Money Market (IMM) date. The return series simulate a strategy of buying and holding the forward contract maturing at the nearest IMM date and rolling to the far contract three business days before the maturity date. Before 1990, when the forward contract data is not available, we use changes in spot exchange rates plus the carry of the currency (difference in local interest rates) for the total return.

We cover 20 developed market currencies, including the G10 (legal tenders of Australia, Eurozone, Canada, Japan, Norway, New Zealand, Sweden, Switzerland, United Kingdom, United States) and 10 legacy European currencies (legal tenders of Belgium, Spain, Finland, France, Germany, Ireland, Italy, Netherlands, Austria, Portugal) before the Euro in 1999. The data set starts in August 1971 when the US severed its link between the value of the US dollar and gold. Panel C of Table B1 reports summary statistics on the currency returns. In addition to large heterogeneity in mean and volatility of currency-pair returns, we also see evidence of currency crashes, which many have claimed are related to carry strategies in currencies (Brunnermeier *et al.*, 2008; Burnside *et al.*, 2010; Koijen *et al.*, 2018).

Commodity futures

We obtain monthly futures prices for 30 commodities starting in February 1877. The source of the data until 1951 is the Annual Report of the Trade and Commerce of the Chicago Board of Trade.[24] Between 1951 and 2012, the futures prices across various contracts are provided by Commodity Systems, Inc. After 2012, the futures prices are from Bloomberg. For base metals and platinum, rolled return series from the S&P, Goldman Sachs, and Bloomberg are used.

The total returns are the sum of spot returns and the "roll-down" on the futures curve. The methodology for computing total returns is as follows. At each month end, we calculate the return on each contract from the previous month end. For each month, we hold the nearest of the contracts whose delivery month is at least 2 months away.[25] For months in which the desired contract does not have a return, we move to the next contract and follow the same procedure until there is a return or until we reach the fifth contract. If there is still no return, we then hold the contract in front of the desired contract. Note that there are days with limit moves in various grains contracts, and we assume that all contiguous limit moves are incorporated into the first move price.[26] This methodology for calculating commodity returns is the same as that used in Moskowitz *et al.* (2012) and Koijen *et al.* (2016). The cross-section of commodities covers energy, base metal, precious metal, agricultural, and livestock sectors, where Panel D of Table B1 reports their summary statistics.

US stock selection

We also supplement the data above with nearly a century's worth of factor return data in US individual equities. The data come from CRSP and begin in July 1926. The US equity data is well known from many studies, so we do not report summary statistics here.

International stock selection

We also examine international individual equities across 21 developed stock markets (those from Frazzini *et al.*, 2020), but note that the longest sample of international individual equity returns begins in 1972 and for our factors (described below) the earliest data point is July 1984. Despite the limited time series, the international equity data provide another asset class to examine the robustness of many of our results.

Cash returns

We use the 3-month local-currency T-bill yield and returns from Global Financial Data as the risk-free cash returns. For a more recent period when LIBOR rates become available, we use the 3-month ICE LIBOR rates or the closest equivalent.

10-Year slope

We compute the cross-sectional average of US, UK, German, and Japanese 10-year government bond yield from Global Financial Data and subtract the average annualized 3-month government bill returns to arrive at a yield curve slope variable or a monetary policy "path" measure. We use the level and 12-month change of this variable to proxy for point-in-time expected future monetary policy stance.

Amihud illiquidity index

We use the illiquidity measure in Amihud (2002).

Baker–Wurgler Sentiment

We use the Sentiment index in Baker and Wurgler (2006) following Equation (2) in the paper. The index is available on http://people.stern.nyu.edu/jwurgler/

Business cycle indicators

We generate three dummy variables for four stages of the economic cycle (slowdown, contraction, expansion, recovery) based on real GDP growth and year-on-year changes in real GDP growth (second derivative on real GDP w.r.t. time). The four stages of economic cycle live in a two-by-two matrix where one axis is the level of growth and the other one changes. The switching from one stage to another is governed by whether the level or changes in growth exceeds one standard deviation of that time series. Once a switch occurs, the economy is deemed to remain in that stage until another switch is triggered.

CPI inflation

We use the monthly year-on-year CPI inflation rate from Global Financial Data, averaged across the US, UK, Germany, and Japan and compute the cross-sectional average across the US, UK, Germany, and Japan to be representative of the world economy. We remove the German inflation data before 1925 to avoid hyperinflation.

EQ volatility

We compute 36-month realized volatility of an equal-weighted developed stock market basket to proxy for equity premium volatility.

GDP growth

We use the quarterly year-on-year GDP growth from Global Financial Data and compute the cross-sectional average across the US, UK, Germany, and Japan to be representative of the world economy.

Geopolitical risk index

We use the Geopolitical risk historical index (GPRH) which starts in 1899 and matches our century-long sample. The data is publicly available on http://www.policyuncertainty.com/gpr.html

Real interest rate

We compute the cross-sectional average of US, UK, German, and Japanese annualized 3-month government bill returns from Global Financial Data and subtract the relevant inflation rate to get real risk-free interest rate. The level and 12-month change of this variable are used to proxy for current monetary policy stance.

Tail risk dummy

We use a dummy variable to proxy for whether stock market realizes tail risk. The variable is 1 if the monthly return of an equal-weighted developed stock market basket is in its lower 5th percentile; 0 otherwise.

Appendix B: Supplemental Tables and Figures

Table B1 Summary statistics on assets.

Country	Mean (annualized)	St.dev (annualized)	Worst 12-month return	Start date
Panel A: Country equity indices				
AR	82.5%	207.1%	−63.7%	1/29/1988
AU	9.2%	14.8%	−41.1%	2/27/1920
BD	37.0%	267.7%	−89.6%	2/27/1920
BG	6.8%	16.3%	−61.5%	1/31/1951
BR	−36.7%	75.3%	−100.0%	7/30/1965
CB	11.3%	31.1%	−69.7%	1/29/1988
CH	11.6%	35.1%	−75.0%	2/29/1996
CL	24.2%	40.5%	−58.1%	2/28/1975
CN	7.7%	14.8%	−45.9%	1/31/1934
DK	9.2%	17.6%	−42.6%	1/30/1970
ES	8.5%	18.9%	−42.1%	4/30/1940
FN	11.1%	24.6%	−58.7%	2/27/1920
FR	11.0%	24.4%	−49.1%	2/27/1920
GR	9.0%	36.8%	−60.8%	1/31/1977
HK	20.4%	33.6%	−70.4%	1/30/1970
HN	12.7%	33.3%	−58.8%	1/31/1991
ID	19.1%	47.6%	−55.6%	1/29/1988
IN	18.8%	30.2%	−47.1%	1/29/1988
IR	6.1%	20.6%	−68.7%	1/29/1988
IS	10.0%	20.0%	−44.7%	2/28/1986
IT	11.0%	28.8%	−50.3%	1/30/1925
JP	11.1%	23.2%	−44.5%	1/31/1921
KO	33.4%	107.5%	−65.9%	2/28/1962
MX	12.2%	24.2%	−42.3%	1/29/1988
MY	11.0%	26.1%	−51.2%	12/29/1972
NL	9.5%	17.4%	−50.2%	1/31/1951
NW	10.4%	23.6%	−50.5%	1/30/1970
NZ	3.7%	19.1%	−55.1%	7/31/1986
OE	6.5%	21.4%	−63.7%	1/30/1970
PH	11.6%	29.5%	−54.2%	1/29/1982
PO	17.5%	53.9%	−71.6%	5/31/1991
PT	2.6%	19.9%	−49.3%	1/29/1988
RS	15.8%	44.1%	−86.3%	1/31/1995
SA	12.2%	21.3%	−40.8%	2/29/1960

Notes: Summary statistics on every asset in our sample, excluding individual stocks, are reported. The in-sample mean log return and standard deviation, worst 12-month log return, and sample start dates are reported for country equity indices (Panel A), fixed income (Panel B), currencies (Panel C), and commodities (Panel D).

Table B1 (*Continued*)

Country	Mean (annualized)	Stdev (annualized)	Worst 12-month return	Start date
SD	9.3%	17.5%	−48.8%	2/27/1920
SG	11.9%	27.5%	−54.7%	1/30/1970
SW	7.7%	16.1%	−39.1%	2/28/1966
TA	14.5%	33.2%	−67.8%	10/30/1987
TH	13.9%	31.5%	−58.9%	5/30/1975
TK	29.7%	68.0%	−70.3%	2/28/1986
UK	7.6%	17.0%	−59.5%	2/27/1920
US	9.9%	19.1%	−63.5%	2/27/1920
VE	338.3%	784.6%	−69.1%	1/29/1988
Panel B: Fixed income				
AU	2.1%	7.6%	−23.3%	2/27/1920
BD	2.6%	6.0%	−18.4%	1/31/1924
BG	1.9%	5.5%	−20.4%	2/27/1920
CL	6.2%	9.9%	−11.0%	10/29/2004
CN	2.2%	5.8%	−23.4%	1/31/1934
CZ	6.9%	9.3%	−24.8%	3/31/1997
DK	2.2%	6.8%	−20.3%	2/27/1920
ES	1.5%	6.6%	−21.0%	2/27/1920
FR	1.9%	6.6%	−24.5%	2/27/1920
HK	2.3%	6.3%	−14.5%	6/30/1993
HN	5.3%	11.8%	−17.7%	7/31/2001
IN	−0.5%	7.2%	−19.7%	2/27/1920
IS	1.8%	3.9%	−10.8%	11/30/1993
IT	1.1%	8.6%	−33.3%	2/27/1920
JP	2.7%	7.3%	−19.0%	2/27/1920
KO	10.7%	20.4%	−25.3%	1/31/1957
MX	2.8%	7.9%	−13.7%	1/31/1995
MY	2.4%	7.2%	−19.3%	1/31/1961
NL	2.3%	7.4%	−21.1%	2/27/1920
PH	9.7%	15.7%	−30.5%	9/30/1996
PO	5.6%	11.6%	−25.1%	4/30/1998
SA	2.4%	10.3%	−36.7%	2/29/1960
SD	1.9%	5.2%	−16.7%	2/27/1920
SG	1.1%	3.9%	−8.2%	12/31/1987
TA	3.1%	5.1%	−11.1%	1/31/1995
TH	4.4%	12.5%	−30.6%	12/31/1979
UK	1.5%	3.9%	−16.1%	1/31/1933
US	2.2%	6.3%	−19.9%	2/27/1920

Table B1 (*Continued*)

Country	Mean (annualized)	St.dev (annualized)	Worst 12-month return	Start date
Panel C: Currencies				
AR	9.1%	19.2%	−42.8%	4/30/1991
AU	2.2%	11.0%	−27.5%	1/31/1972
BD	1.3%	10.8%	−30.7%	1/29/1971
BG	3.0%	12.0%	−34.6%	9/30/1971
BR	8.9%	17.8%	−30.2%	7/29/1994
BU	0.3%	10.6%	−21.6%	3/31/2005
CB	2.9%	11.8%	−34.2%	2/28/1992
CH	1.5%	2.8%	−5.9%	2/29/1996
CL	−7.3%	11.8%	−49.1%	6/30/1982
CN	0.3%	6.7%	−21.6%	1/31/1972
CZ	2.6%	11.4%	−22.7%	4/30/1993
ES	2.1%	10.9%	−29.1%	9/30/1971
FN	2.8%	11.1%	−28.3%	9/30/1971
FR	2.8%	11.1%	−30.0%	9/30/1971
GR	4.0%	11.5%	−30.7%	9/30/1971
HK	−5.0%	3.6%	−32.0%	12/30/1977
HN	3.5%	12.2%	−22.8%	8/31/1989
ID	2.4%	19.9%	−72.5%	12/31/1981
IN	1.0%	7.0%	−18.8%	1/29/1993
IR	2.5%	11.6%	−30.7%	9/30/1971
IS	0.4%	7.9%	−22.9%	1/31/1986
IT	3.6%	10.9%	−27.3%	9/30/1971
JP	1.3%	11.1%	−28.3%	1/29/1971
KO	−1.1%	11.4%	−46.6%	1/31/1980
MX	6.3%	23.7%	−64.6%	1/31/1980
MY	−2.9%	7.8%	−43.3%	11/28/1986
NL	2.3%	11.7%	−31.9%	9/30/1971
NW	1.0%	11.0%	−25.0%	12/30/1977
NZ	3.0%	12.0%	−32.6%	12/30/1977
OE	2.6%	11.7%	−30.0%	9/30/1971
PH	−1.2%	9.9%	−38.1%	1/31/1980
PO	4.4%	12.3%	−32.7%	1/31/1992
PT	2.9%	11.7%	−30.7%	9/30/1971
RS	6.7%	15.7%	−45.0%	9/30/1994
SA	−3.5%	14.2%	−39.1%	11/30/1972
SD	0.2%	10.8%	−30.2%	12/30/1977
SG	0.5%	5.4%	−18.0%	3/31/1986
SW	1.8%	11.8%	−28.7%	1/29/1971

Table B1 (*Continued*)

Country	Mean (annualized)	Stdev (annualized)	Worst 12-month return	Start date
TA	−1.6%	5.6%	−18.7%	1/31/1985
TH	1.2%	9.5%	−46.5%	5/31/1988
TK	8.1%	16.2%	−38.9%	9/30/1996
UK	1.2%	9.9%	−27.1%	1/29/1971
US	0.0%	0.0%	0.0%	10/31/1967
VE	−0.6%	22.8%	−49.9%	1/31/1989
Commodity	Mean (annualized)	Stdev (annualized)	Worst 12-month return	Start date
Panel D: Commodities				
ALUMINUM	−0.2%	18.9%	−57.8%	7/31/1992
BRENTOIL	15.0%	33.3%	−56.4%	7/29/1988
CATTLE	5.6%	16.4%	−35.7%	12/31/1964
COCOA	9.9%	32.0%	−51.4%	1/31/1966
COFFEE	10.9%	39.5%	−57.5%	9/29/1972
COPPER	12.1%	24.8%	−52.7%	9/30/1993
CORN	5.3%	26.2%	−61.5%	2/28/1877
COTTON	6.0%	24.7%	−56.4%	2/27/1925
CRUDE	14.4%	39.9%	−66.6%	4/29/1983
FEEDERCATTLE	4.6%	16.7%	−50.3%	12/31/1971
GASOIL	11.1%	31.5%	−57.3%	5/29/1981
GOLD	3.6%	19.1%	−43.0%	2/28/1975
HEATOIL	12.8%	33.9%	−57.1%	12/29/1978
HOGS	5.5%	26.3%	−51.2%	3/31/1966
KANSASWHEAT	3.7%	26.5%	−53.3%	6/30/1966
LARD	1.3%	24.5%	−55.8%	2/28/1877
LEAD	10.0%	27.8%	−63.2%	2/28/1995
NATGAS	1.9%	50.4%	−75.3%	5/31/1990
NICKEL	14.6%	35.0%	−64.4%	10/31/1994
OATS	6.8%	31.2%	−60.4%	2/28/1877
PORK	7.9%	30.4%	−51.1%	2/28/1877
SHORTRIBS	12.4%	24.8%	−53.8%	2/28/1885
SILVER	8.8%	32.6%	−60.0%	7/31/1963
SOYBEANS	10.6%	26.3%	−40.1%	2/26/1937
SOYMEAL	13.4%	30.5%	−60.9%	9/28/1951
SOYOIL	10.2%	29.8%	−53.0%	8/31/1950
SUGAR	12.8%	44.8%	−69.6%	1/31/1966
UNLEADED	20.8%	37.2%	−52.3%	1/31/1985
WHEAT	4.3%	25.0%	−53.7%	2/28/1877
ZINC	4.4%	24.7%	−57.4%	10/30/1992

Table B2 Structural break tests for factor returns and variances.

Asset Class	Value	Momentum	Carry	Defensive	Multifactor	Market
Panel A: Chow test of structural break in returns						
By decade						
Commodities	0.016	0.022	0.004	n.a.	0.005	0.029
Equity country indices	0.098	0.025	0.018	0.003	**0.000**	0.005
Fixed income	0.370	0.826	0.160	0.120	0.585	**0.000**
Currencies	0.163	0.649	0.575	n.a.	0.020	0.476
U.S. stocks	0.046	0.137	n.a.	0.193	0.081	0.289
International stocks	0.009	0.634	n.a.	0.083	0.031	0.706
Multi-asset	0.082	0.238	0.002	0.006	0.082	0.082
By 20-year periods						
Commodities	0.006	0.022	0.002	n.a.	0.001	0.018
Equity country indices	0.205	0.032	0.024	0.170	0.002	**0.001**
Fixed income	0.381	0.863	0.019	0.058	0.618	**0.000**
Currencies	0.438	0.502	0.556	n.a.	0.041	0.674
U.S. stocks	0.799	0.441	n.a.	0.236	0.479	0.366
International stocks	0.567	0.592	n.a.	0.012	0.008	0.913
Multi-asset	0.204	0.482	0.002	0.008	0.030	0.074
Panel B: Chow test of structural break in variance						
By decade						
Commodities	**0.000**	**0.000**	**0.000**	n.a.	**0.000**	**0.000**
Equity country indices	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Fixed income	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Currencies	**0.000**	0.105	**0.000**	n.a.	**0.000**	**0.000**
U.S. stocks	**0.000**	**0.000**	n.a.	**0.000**	**0.000**	**0.000**
International stocks	**0.000**	0.004	n.a.	**0.000**	**0.000**	0.027
Multi-asset	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
By 20-year periods						
Commodities	0.030	0.123	0.157	n.a.	0.061	**0.000**
Equity country indices	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Fixed income	**0.000**	**0.000**	0.002	**0.000**	**0.000**	**0.000**
Currencies	**0.000**	0.117	**0.000**	n.a.	**0.000**	0.062
U.S. stocks	0.006	**0.000**	n.a.	**0.000**	**0.000**	**0.000**
International stocks	0.663	0.163	n.a.	0.690	0.360	0.283
Multi-asset	0.019	0.003	**0.000**	**0.000**	**0.000**	**0.000**
Panel C: Sup-Wald test of structural breaks in returns						
Commodities			**1953-8-31**	n.a.	**1967-6-30**	
Equity country indices		2000-3-31			2005-1-31	
Fixed income			1986-1-31			**1981-10-30**

Table B2 (*Continued*)

Asset Class	Value	Momentum	Carry	Defensive	Multifactor	Market
Currencies				n.a.		
U.S. stocks			n.a.	1975-4-30	2005-10-31	
International stocks	**2010-5-31**		n.a.	1998-12-31	2000-9-29	
Multi-asset			1945-1-31	1980-12-31	2006-5-31	
Panel D: Sup-Wald test of structural breaks in variances						
Commodities				n.a.	2002-9-30	**1951-7-31**
Equity country indices	**2002-5-31**	**2003-8-29**	**2003-7-31**	**2004-4-30**	**2003-10-31**	**1971-12-31**
Fixed income	**1996-5-31**	**1996-3-29**	**1995-8-31**	**1995-9-29**	**1994-9-30**	**1973-12-31**
Currencies	**1998-11-30**		**1985-12-31**	n.a.	**1987-9-30**	**2012-7-31**
U.S. stocks	1943-5-31		n.a.	**1997-1-31**	**2005-3-31**	**1941-8-29**
International stocks	**2001-10-31**	**2009-9-30**	n.a.	**2010-1-29**	**2009-5-29**	2012-7-31
Multi-asset	**1988-2-29**	**1947-8-29**	**1996-4-30**	**2005-1-31**	**2002-10-31**	**1979-10-31**

Notes: Panel A reports structural break tests for mean returns of the factors in each asset class and the market portfolio in each asset class using the Chow test with decade breakpoints and with 20-year breakpoints separately. Panel A reports p-values from the Chow test assuming heteroskedasticity, where shaded numbers indicate rejection of the null with less than 5% significance assuming independent tests and bold numbers indicate rejection with less than 5% significance taking into account the multiple comparisons being made using the Bonferroni correction. Panel B repeats the Chow tests for structural changes in the variance of each factor by asset class. Panels C and D report results for structural break tests using the sup-Wald test (Andrews, 1993, 2003), where the first breakpoint dates that pass the 10% significance level are identified in the table. Bold dates indicate multiple breakpoints are found, non-bolded implies only one reliable breakpoint found, and no date indicates no significant breakpoints are found. Panel C reports break points for mean returns and Panel D for the variance of returns.

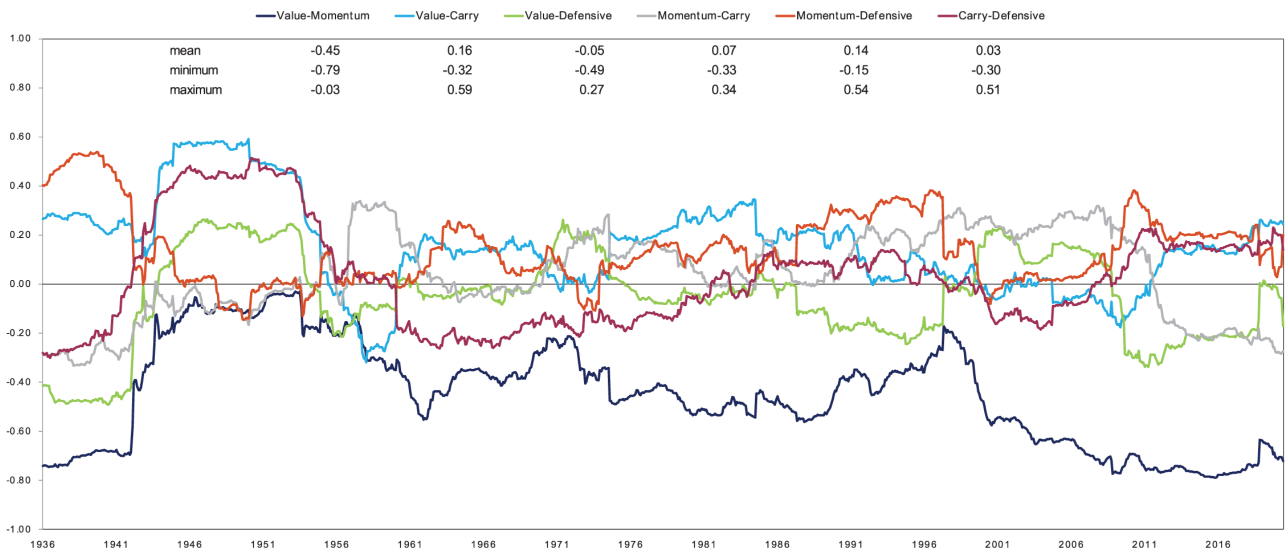


Figure B1 Time-varying diversification.

Figure plots the pairwise correlations between factors (across all asset classes) over time using rolling monthly return data over the prior 5 years to estimate the correlations at each point in time. The mean, minimum, and maximum realized correlations are also reported on the graph.

Notes

[1] The list of dynamic asset pricing models is numerous. Starting with the intertemporal models of Merton (1973) and Breeden (1979), a variety of mechanisms that generate dynamics have been proposed, including preferences over consumption, such as habit (e.g., Constantinides, 1990; Campbell and Cochrane, 1999), long-run risks (Epstein and Zin, 1991; Bansal and Yaron, 2004), ambiguity aversion (Epstein and Schneider, 2008), rare disasters (Barro and Ursua, 2012; Wachter, 2013), production-based models (Cochrane, 1996; Belo, 2010), prospect theory (Barberis *et al*., 2001), and extrapolative beliefs (Barberis *et al*., 2015).

[2] A recent paper by Jensen *et al*. (2021) refutes some of the results in these papers showing more statistical support for factor studies including applying a Bayesian approach to evaluating out-of-sample evidence.

[3] See Fama and French (2012), Gorton *et al*. (2013), Asness *et al*. (2013, 2015), Bhojraj and Swaminathan (2005), Frazzini and Pedersen (2013), Koijen *et al*. (2018), Brooks and Moskowitz (2018), and Baltussen *et al*. (2019). Most return histories start in the 1960s or later, except for US stock returns available since the 1920s. Some studies focus on a long history of one factor—e.g., Geczy and Samonov (2016) who study a 200-year track record on momentum—and study a narrower set of questions. Baltussen *et al*. (2019) is closest to our sample as it uses a 200-year history of several factor premia, some of which are different factors and different asset classes, but focuses mainly on statistical robustness. While our findings on factor premia existence complement their statistical robustness results, our focus is primarily on conditional variation and predictability over time.

[4] The factor return data is available here: www.aqr.com/Insights/Datasets and will be updated through time.

[5] We use CAPE for country equity indices instead of aggregate book-to-price since book values are not available as far back and since Campbell and Shiller (1998) find CAPE to be a better description of equity market valuation.

[6] While alternate measures (such as survey forecasts) provide a measure of inflation expectations that is forward looking, these forecasts are only available over the recent time period (beginning 1990). Using historical 3-year moving averages of inflation has the benefit of being available consistently over our longer sample period and, as Asness *et al*. (2013) show, delivers similar results over the same time period.

[7] PPP exchange rates are equilibrium exchange rates that make a basket of goods equally expensive in two countries. The currency value portfolio buys currencies whose nominal exchange rate is lower than the PPP exchange rate and sells currencies whose nominal exchange rate is higher than PPP.

[8] Skipping a month in forming momentum, as in Asness (1994), appears important for individual stocks (likely more so in our very early sample period where liquidity may be an issue), but is less important, if unnecessary, for other asset classes. Nevertheless, we use the same definition of momentum across all asset classes.

[9] For USD, EUR, JPY, GBP, CHF, and all legacy European currencies, we use the 3-month ICE LIBOR daily rate. For most other currencies, we find the local equivalent 3-month interbank offered rate that uses a methodology similar to that of ICE LIBOR. For example, we use the Prague Interbank Offered Rate for CZK. If there is no obvious LIBOR equivalent, we use 3-month bank bill/CD returns, deposit rates, or swap rates as a substitute.

[10] Trading costs and other implementation frictions tend to be larger in individual equities, so net of cost Sharpe ratios may be more similar across asset classes, though we do not have the data to test this conjecture explicitly.

[11] While it is possible that some traders knew of these factors before their academic discovery, it was not publicized or widely disseminated relative to the period after its publication. Moreover, our pre-sample period pre-dates the beginning of the original sample period, which is often several decades before the publication date. Over this distant period, it seems reasonable to assume that little arbitrage activity was taking place in the factor.

[12] A caveat to the pre-sample data is that the data may be of poorer quality.

[13] For value, this is Fama and French (1992), whose original sample period is July 1963 to July 1990. For momentum, it is Jegadeesh and Titman (1993), whose original sample period is January 1964 to December 1989. For carry, we use Meese and Rogoff (1983) and Fama (1984), who document a "forward premium puzzle" in currencies that is the basis for the carry trade, and whose original sample period is 1973 to 1982. For defensive, we use the original sample period of Frazzini and Pedersen (2013) from 1960 to 2009, since we focus on low beta strategies and their betting-against-beta (BAB) factor.

[14] Accominotti et al. (2019) find positive carry and momentum returns in currencies in the 1920s and 1930s

that precede the fixed rate regime of Bretton Woods, providing additional out-of-sample evidence.

15 In fact, the genesis of Asness *et al*. (2013) came from a practitioner (Asness) implementing value and momentum strategies across many asset classes and markets shortly after their discovery in the mid-1990s, but nearly 20 years prior to the publication of Asness *et al*. (2013).

16 In addition, as Daniel and Moskowitz (2016) show, momentum crashes during the best global market months, where value appears to be a particularly valuable hedge to those momentum crashes with its −0.66 correlation during these times.

17 For example, dividend yield (Fama and French, 1988) or CAPE (cyclically-adjusted P/E ratio in Campbell and Shiller, 1998) for the aggregate US equity market, valuation spreads for the value factor (Asness *et al*., 2000; Cohen *et al*., 2003), and forward rates for bonds (Fama and Bliss, 1987).

18 See Brooks and Moskowitz (2018) in global bond markets, and Stambaugh *et al*. (2012), Greenwood and Hanson (2012), Akbas *et al*. (2015), Gupta and Kelly (2019), and Arnott *et al*. (2019) for multiple factors in US equities.

19 See Baba *et al*. (2019) and Asness *et al*. (2018).

20 See Asness *et al*. (2000, 2017, 2018), Vuolteenaho (2002), Cohen *et al*. (2003), and Lochstoer and Tetlock (2016).

21 Instead of the difference between the longs and shorts, the literature often takes the ratio or log difference, which is less contaminated by rising price levels over time that affect aggregate valuations. On the other hand, ratios can be problematic if the denominator is small or negative. We use ratios of valuations for equity factors (following the literature and because small or negative denominators are not a problem here), and use differences in valuation metrics for factors in other asset classes, where small or negative valuations can frequently occur. Results are not sensitive to using either method.

22 Restricting our analysis to roughly the same sample period as these other studies, and focusing exclusively on US equities, we find larger factor momentum timing profits, consistent with these other studies. We also focus on the past 12-month factor momentum return signal, while the largest signal found in Arnott *et al*. (2019) and Gupta and Kelly (2019) is a shorter-term 1-month return signal.

23 We also examine a new timing methodology suggested by Haddad *et al*. (2019) applied to our factors and asset classes and to all of our timing signals. The out-of-sample timing results are meager. One reason we get poorer out-of-sample results than Haddad *et al*. (2019) is that our factors across asset classes have a weaker covariance structure than the 50 equity-only factors Haddad *et al*. (2019) study, which is an important feature of their methodology.

24 Commodity returns are missing from 1944 to 1949 due to a minimum asset requirement not being met as there are too few commodities with reliable data during this period.

25 For example, we hold an April contract through the end of February. An exception is Brent oil, whose delivery month needs to be at least 3 months away, i.e. we hold the April contract through the end of January. This methodology is chosen to coincide with the procedure employed by the popular Goldman Sachs Commodity Index.

26 For limit day periods, we incorporate all the limit day returns into the first limit day following Roll (1984). Limit days are determined by whether on that day (i) the maximum price shift across contracts of the same commodity is a round amount (before closing prices are available, the largest positive shift from high price and the largest negative shift from low price are used) (ii) two or more contracts move by this amount, (iii) if maximum price shift is from the front contract and does not meet above conditions, maximum shift of the other contracts meets above conditions (since sometimes the front contract is not subject to limits if it is considered "spot"), and (iv) this shift is equal to or higher than the official price limit set by the exchange (when available).

Acknowledgments

We thank Ron Alquist, Cliff Asness, Jacob Boudoukh, Jeff Dunn, Stefano Giglio, Sarah Jiang, Michael Katz, Bryan Kelly, John Liew, Hanno Lustig, Lasse Pedersen, and seminar participants at AQR and the Swedish Institute for Financial Research for their valuable comments and suggestions. Moskowitz thanks the International Finance Center at Yale University for financial support. AQR Capital Management is a global investment management firm, which may

or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. The century of factor returns used in this paper will be made available on AQR's data library webpage and will be updated over time.

References

Acharya, V. and Pedersen, L. H. (2005). "Asset Pricing with Liquidity Risk," *Journal of Financial Economics* **77**, 375–410.

Accominotti, O., Cen, J., Chambers. D., and Marsh I. (2019). "Currency Regimes and the Carry Trade," *Journal of Financial and Quantitative Analysis* **54**(5), 2233–2260.

Akbas, F., Armstrong, W. J., Sorescu, S., and Subrahmanyam, A. (2015). "Smart Money, Dumb Money, and Capital Market Anomalies," *Journal of Financial Economics* **118**(2), 355–382.

Alquist, R., Jiang, S., and Moskowitz, T. J. (2019). "Crowding," Working Paper, AQR Capital.

Amihud, Y. (2002). "Illiquidity and Stock Returns: Cross-section and Time Series Effects," *Journal of Financial Markets* **5**, 31–56.

Arnott, R. D., Clements, M., Kalesnik, V., and Linnainmaa, J. (2019). "Factor Momentum," Working paper, Research Affiliates.

Asness, C. S. (1994). "Variables That Explain Stock Returns," Ph.D. dissertation, University of Chicago.

Asness, C., Chandra, S., Ilmanen, A., and Israel, R. (2017). "Contrarian Factor Timing is Deceptively Difficult," *The Journal of Portfolio Management* **43**(5), 72–87.

Asness, C., Friedman, J. A., Krail R. J., and Liew, J. M. (2000). "Style Timing: Value versus Growth," *Journal of Portfolio Management* **26**(3), 50–60.

Asness, C., Ilmanen, A., Israel, R., and Moskowitz, T. (2015). "Investing with Style," *Journal of Investment Management* **13**(1), 27–63.

Asness, C., Ilmanen, A., and Maloney, T. (2017). "Market Timing: Sin a Little," *Journal of Investment Management* **15**(3), 23–40.

Asness, C., Liew, J., Pedersen, L., and Thapar, A. (2021). "Deep Value," *The Journal of Portfolio Management* **47**(4), 11–40.

Asness, C., Moskowitz, T. J., and Pedersen, L. (2013). "Value and Momentum Everywhere," *Journal of Finance* **68**(3), 929–985.

Baba, F., Boons, M., and Tamoni, A. (2019). "Value Return Predictability across Asset Classes and Commonalities in Risk Premia," SSRN Working Paper.

Baker, S., Bloom, N., and Davis, S. (2016). "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics* **131**(4) (July), 1593–1636.

Baker, M. and Wurgler, J. (2006). "Investor Sentiment and the Cross-section of Stock Returns," *The Journal of Finance* **61**(4), 1645–1680.

Baltussen, G., Swinkels, L., and Van Vliet, P. (2019). "Global Factors Premiums," Robeco Working Paper.

Bansal, R., and Yaron, A. (2004). "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance* **59**(4), 1481–1509.

Barberis, N., Greenwood, R., Jin, L., and Shleifer, A. (2015). "X-CAPM: An Extrapolative Capital Asset Pricing Model," *Journal of Financial Economics* **115**(1), 1–24.

Barro, R. and Ursua, J. F. (2012). "Rare Macroeconomic Disasters", *Annual Review of Economics* 83–109.

Belo, F. (2010). "Production-Based Measures of Risk for Asset Pricing," *Journal of Monetary Economics* **57**, 146–163.

Berk, J., Green, R., and Naik, V. (1999). "Optimal Investment, Growth Options, and Security Returns," *The Journal of Finance* **54**, 1153–1607.

Boudoukh, S. and Swaminathan, B. (2006). "Macromomentum: Returns Predictability in International Equity Indices," *The Journal of Business* **79**(1), 429–451.

Boudoukh, J., Israel, R., and Richardson, M. (2020). "The Statistics of Long-Horizon Regressions," AQR Capital Working Paper.

Breeden, D. (1979). "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics* **7**, 265–2966.

Brooks, J. and Moskowitz, T. J. (2018). "Yield Curve Premia," AQR Capital and Yale University Working Paper.

Brunnermeier, M. K., Nagel, S., and Pedersen, L. H. (2008). "Carry Trades and Currency Crashes," *NBER Macroeconomics Annual* **23**, 313–348.

Caldora, D. and Iacoviello, M. (2018). "Measuring Geopolitical Risk," Federal Reserve Bank Working Paper.

Campbell, J. and Cochrane, J. (1999). "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock

Market Behavior," *Journal of Political Economy* **107**(2), 205–251.

Campbell, J. and Shiller, R. (1988). "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies* **1**(3), 195–228.

Campbell, J. and Shiller, R. (1998). "Valuation Ratios and the Long-Run Stock Market Outlook," *Journal of Portfolio Management* **24**(2), 11–26.

Campbell, J. Y. and Thompson, S. B. (2007). "Predicting Excess Stock Returns Out-of-Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies* **21**(4), 1509–1531.

Carlson, M., Fisher, A., and Giammarino, R. (2004). "Corporate Investment and Asset Price Dynamics: Implications for the Cross Section of Returns," *Journal of Finance* **59**, 2577–2603.

Chen, N. F., Roll, R., and Ross, S. (1986). "Economic Forces and the Stock Market," Journal of Business, **59**(3), 383–403.

Chordia, T. and Shivakumar, L. (2005). "Momentum, Business Cycle, and Time-Varying Expected Returns," *The Journal of Finance* **57**(2), 985–1019.

Cochrane, J. H. (1991). "Production-Based Asset Pricing and the Link between Stock Returns and Economic Fluctuations," *The Journal of Finance* **46**, 209–237.

Cochrane, J. H. (1996). "A Cross-Sectional Test of an Investment-Based Asset Pricing Model," *Journal of Political Economy* **104**, 572–621.

Cohen, R. B., Polk, C., and Vuolteenaho, T. (2003). "The Value Spread," *The Journal of Finance* **58**(2), 609–641.

Constantinides, G. (1990). "Habit Formation: A Resolution of the Equity Premium Puzzle," *The Journal of Political Economy* **98**(3), 519–543.

Cooper, I. and Priestley, R. (2011). "Real Investment and Risk Dynamics," *Journal of Financial Economics* **101**, 182–205.

Daniel, K. and Moskowitz, T. J. (2016). "Momentum Crashes," *Journal of Financial Economics* **122**(2), 221–247.

DeBondt, W. and Thaler, R. (1985). "Does the Stock Market Overreact?" *The Journal of Finance* **40**(3), 793–805.

Ehsani, S. and Linnainmaa, J. T. (2019). "Factor Momentum and the Momentum Factor." SSRN Working Paper.

Epstein, L. and Schneider, M. (2008). "Ambiguity, Information Quality, and Asset Pricing," *Journal of Finance* **63**(1), 197–228.

Epstein, L. and Zin, S. (1991). "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis," *The Journal of Political Economy* **99**(2), 263–286.

Fama, E. F. (1984). "Forward and Spot Exchange Rates," *Journal of Monetary Economics* **14**, 319–338.

Fama, E. and Bliss, R. (1987). "The Information in Long-Maturity Forward Rates," *The American Economic Review*, 680–692.

Fama, E. and French, K. (1988). "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics* **22**(1), 3–25.

Fama, E. F. and K. R. French (1992). "The Cross-Section of Expected Stock Returns," *Journal of Finance* **47**, 427–465.

Fama, E. F. and French, K. R. (1993). "Common Risk Factors in the Returns on Stock and Bonds" *Journal of Financial Economics* **33**, 3–56.

Fama, E. F. and French, K. R. (1996). "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance* **51**, 55–84.

Fama, E. F. and French, K. R. (2012). "Size, Value, and Momentum in International Stock Returns," *Journal of Financial Economics* **105**(3), 457–472.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2020). "Trading Costs of Asset Pricing Anomalies," Working paper Yale University and AQR Capital.

Frazzini, A. and Pedersen, L. H. (2013). "Betting against Beta," *Journal of Financial Economics* **111**(1), 1–25.

Gabaix, X. (2012). "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzle in Macro Finance," *Quarterly Journal of Economics* **127**(2), 645–700.

Geczy, C. G. and Samonov, M. (2016). "Two Centuries of Price Return Momentum," *Financial Analysts Journal* **72**(5).

Gomes, J. F., Kogan, L., and Zhang, L. (2003). "Equilibrium Cross Section of Returns," *Journal of Political Economy* **111**, 693–732.

Gormsen, N. and Lazarus, E. (2019). "Duration-Driven Returns," SSRN Working Paper.

Gorton, G., Hayashi, F., and Rouwenhorst, G. (2013). "The Fundamentals of Commodity Futures Returns,"*Review of Finance*, European Finance Association **17**(1), 35–105.

Greenwald, B., Lettau, M., and Ludvigson, S. (2014). "Origins of Stock Market Fluctuations," National Bureau of Economic Research Working Paper No. 19818.

Greenwood, R. and Hanson, S. G. (2012). "Share Issuance and Factor Timing," *The Journal of Finance* **67**(2), 761–798.

Griffin, J., Ji, X., and Martin, J. S. (2003). "Momentum Investing and Business Cycle Risk: Evidence from Pole to Pole," **58**(6), 2515–2547.

Gupta, T. and Kelly, B. (2019). "Factor Momentum Everywhere," *Journal of Portfolio Management* **45**(3), 13–36.

Hansen, L. P. and Richard, S. F. (1987). "The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models," *Econometrica* **55**(3), 587–613.

Haddad, V., Kozak, S., and Santosh, S. (2019). "Factor Timing," Working Paper UCLA.

Hansen, L. P., Heaton, J., and Li, N. (2008). "Consumption Strikes Back? Measuring Long-Run Risk," *Journal of Political Economy* **116**(2), 260–302.

Hansen, L. P. and Jagannathan, R. (1991). "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy* **99**(2), 225–262.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). "And the Cross-section of Expected Returns," *Review of Financial Studies* **29**(1), 5–68.

Herskovic, B., Moreira, A., and Muir, T. (2019). "Hedging Risk Factors," UCLA Working Paper.

Hodges, P., Hogan, K., Peterson, J. R., and Ang, A. (2017). "Factor Timing with Cross-sectional and Time-Series Predictors," *The Journal of Portfolio Management* **44**(1), 30–43.

Hou, K., Xue, C., and Zhang, L. (2015). "Digesting Anomalies: An Investment Approach", *Review of Financial Studies* **28**, 650–705.

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

Jegadeesh, N. and Titman, S. (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance* **48**, 65–91.

Jensen, T. I., Kelly, B., and L. H. Pedersen (2021). "Is There a Replication Crisis in Finance?" NBER Working Paper 28432.

Jiang, H. and Kelly, B. (2014). "Tail Risk and Asset Prices," *The Review of Financial Studies* **27**(10), 2841–2871.

Johnson, T. (2002). "Rational Momentum Effects," *The Journal of Finance* **57**(2) (April 2002), 585–608.

Koijen, R., Moskowitz, T. J., Pedersen, L., and Vrugt, E. (2018). "Carry," *Journal of Financial Economics* **127**(2), 197–225.

Lettau, M., Maggiori, M., and Weber, M. (2014). "Conditional Risk Premia in Currency Markets and Other Asset Classes," *Journal of Financial Economics* **114**(2), 197–225.

Lettau, M. and Ludvigson, S. (2001). "Resurrecting the (C)CAPM: A Cross-sectional Test When Risk Premia are Time-Varying" *Journal of Political Economy* **109**(6), 1238–1287.

Lettau, M. and Ludvigson, S. (2009). "Euler Equation Errors," *Review of Economic Dynamics* **12**(2), 255–283.

Lettau, M. and Wachter, J. (2007). "Why is Long-horizon Equity Less Risky? A Duration-Based Explanation of the Value Premium," *Journal of Finance* **LXII**(1), 55–92.

Lewellen, J., Nagel, S., and Shanken, J. (2010). "A Skeptical Appraisal of Asset Pricing Tests," *Journal of Financial Economics* **96**, 175–194.

Li, D. and Zhang, L. (2010). "Does q-Theory with Investment Frictions Explain Anomalies in the Cross Section of Returns? *Journal of Financial Economics* **98**(2010), 297–314.

Li, E. X. N., Livdan, D., and Zhang, L. (2009). "Anomalies," *Review of Financial Studies* **22**(2009), 4301–4334.

Linnainmaa, J. and Roberts, M. (2018). "The History of the Cross-section of Stock Returns," *Review of Financial Studies* **31**(7), 2606–2649.

Liu, L. X. L. Whited, T. M., and Zhang, L. (2009). "Investment-Based Expected Stock Returns," *Journal of Political Economy* **117**, 1105–1139.

Liu, L. X. L. and Zhang, L. (2014). "A neoclassical interpretation of momentum," *Journal of Monetary Economics* **67**, 109–128.

Lochstoer, L. A. and Tetlock, P. C. (2016). "What Drives Anomaly Returns?" Columbia University Working Paper.

Lou, D. and Polk, C. (2021). "Comomentum: Inferring Arbitrage Activity from Return Correlations," London School of Economics working paper.

Malloy, C., Moskowitz, T. J., and Jorgensen, A. V. (2009). "Long-Run Stock Holder Consumption Risk and Asset Returns," *The Journal of Finance* **LXIV**(6), 2427–2480.

McLean, R. D. and Pontiff, J. (2016). "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance* **71**, 5–32.

Meese, R. A. and Rogoff, K. (1983). "Empirical Exchange Rate Models of the Seventies: Do They Fit out of Sample?" *Journal of International Economics* **14**, 3–24.

Merton, R. C. (1973). "An Intertemporal Capital Asset Pricing Model," *Econometrica: Journal of the Econometric Society*, 867–887.

Moreira, A. and Muir, T. (2017). "Volatility-Managed Portfolios," *The Journal of Finance* **72**(4), 1611–1644.
Moskowitz, T., Yao Hua Ooi, J., and Pedersen, L. (2012). "Time Series Momentum," *Journal of Financial Economics* **104**(2), 228–250.
Parker, J. A. and Julliard, C. (2005). "Consumption Risk and the Cross Section of Expected Returns," *Journal of Political Economy* **113**, 185–222.
Pastor, L. and Stambaugh, R. F. (2003). "Liquidity Risk and Expected Returns," *Journal of Political Economy* **111**(3), 642–685.
Roll, R. (1984). "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," *Journal of Finance* **39**(4), 1127–1139.
Sadka, R. (2006). "Momentum and Post-Earnings-Announcement Drift Anomalies: The Role of Liquidity Risk," *Journal of Financial Economics* **80**, 309–349.
Schwert, G. W. (2003). "Anomalies and Market Efficiency," In: Constantinides, G. M., M. Harris, and R. M. Stulz, eds., pp. 939–974. Elsevier, Amsterdam: *Handbook of the Economics of Finance*.
Stambaugh, R. F. (1999). "Predictive Regressions," *Journal of Financial Economics* **54**, 375–421.
Stambaugh, R. F., Yu, J., and Yuan, Y. (2012). "The Short of it: Investor Sentiment and Anomalies," *Journal of Financial Economics* **104**(2), 288–302.
Tsai, J. and Wachter, J. A. (2015). "Disaster Risk and its Implications for Asset Pricing," *Annual Review of Financial Economics* **7**, 219–252.
Vuolteenaho, T. (2002). "What Drives Firm-Level Stock Returns?" *The Journal of Finance* **57**(1), 233–264.
Wachter, J. (2013). "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" *Journal of Finance* **68**, 987–1035.
Xing, Y. (2008). "Interpreting the Value Effect through the Q-Theory: An Empirical Investigation," *Review of Financial Studies* **21**, 1767–1795.
Zhang, L. (2005). "The Value Premium," *Journal of Finance* **60**, 67–103.

Keywords: Factor investing, multi-factor and multi-asset class portfolios, factor timing, tactical asset allocation, diversification