The Review of Financial Studies

A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction

Amit Goyal
Swiss Finance Institute, Switzerland

Ivo Welch
UCLA Anderson, USA

Athanasse Zafirov

Our paper reexamines whether 29 variables from 26 papers published after Goyal and Welch 2008, as well as the original 17 variables, were useful in predicting the equity premium in-sample and out-of-sample as of the end of 2021. Our samples include the original periods in which these variables were identified, but end later. More than one-third of these new variables no longer have empirical significance even in-sample. Of those that do, half have poor out-of-sample performance. A small number of variables still perform reasonably well both in-sample and out-of-sample. (*JEL* G10, G12, G14, G17)

Received: February 13, 2022; Editorial decision: January 13, 2024
Editor: Ralph Koijen
Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

Since Goyal and Welch 2008, henceforth GW, a large number of papers predicting the equity premium have been published in top finance journals. Most have been well aware of the GW critique. Many of these papers have further suggested that their variables are built on strong theoretical foundations, presumably increasing confidence that they would be stable forward-looking. It thus seems that academic finance has conquered the problem of predicting time-varying future equity premiums.

Yet, the last two decades have also offered a fascinating new sample with plenty of opportunities for predictive variables to prove their mettle.

We apologize for not thanking authors and seminar participants individually. The referees and editor (Ralph Kojien) were unusually helpful. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Amit Goyal, amit.goyal@unil.ch.

The Review of Financial Studies 37 (2024) 3490–3557
© The Authors 2024. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
https://doi.org/10.1093/rfs/hhae044 Advance Access publication September 4, 2024

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Interrupting a long economic boom, there were three economic recessions, one in the early 2000s (with the dot-com end and Sept. 11), one in 2008 (the Great Recession), and one in 2020 (the COVID-19 recession); and despite a long bull market, there were two major bear markets, one in 2000–2002 and the other in 2008 (plus a minor one in 2018).

It does not indict papers or authors if findings no longer hold after publication. Except for tautologies, empirical external validity in the social sciences can never be taken for granted. This is even more the case when investors could actively attempt to profit from the findings in these papers (McLean and Pontiff 2016). All empirical social science research deserves skeptical reexamination.

Our own paper thus poses two simple questions: First, how reliably have the new predictors performed? As of early 2022, which of the findings of this vibrant literature are still up-to-date? Or do at least some of the predictors—for all their innovativeness, foundations, and empirical support—no longer warrant the same trust that they garnered at publication? Second, how robust are the predictors to various alternative specifications?

Our paper therefore reexamines 29 variables proposed in 26 prominent recently published papers (listed in Table 1), for which we could relatively easily reconstruct and/or obtain the predictive variable. The data in these original papers ended between 2000 and 2017. We could replicate and confirm the principal findings for almost all predictors, using a simple predictive framework based on OLS in-sample (IS) univariate forecasting regressions.

We then extend the samples forward, ending with financial market returns in December 2021 and typically about 10 extra years of data. Because our paper includes the data that the authors themselves had originally used to discover and validate their theories, our tests are neither harsh nor independent. Most of what the predictors had to do in the few added years was not to "screw up" badly. The original results should still hold. If the theories are stable (and the data is representative), the statistical significance of their findings should creep upward (on average), with further power provided by more data.

Yet, we find that more than one-third of our variables have already lost their predictive power. We replicated 29 variables in Table 2, and added July to June intervals for two annual predictors with good January to December performance, giving us 31 variables. One predictor (semiannual) could not be replicated even in the author's original sample. With data extended to 2021, 13 variables had no in-sample predictive ability. To put this into perspective, simulations (Internet Appendix IA.I) suggest that such poor extended in-sample performance seems more consistent with a scenario in which the 29 variables were all spurious to begin with, rather than with one in which there was a modest stable association between predictors and market premiums.

Harvey, Liu, and Zhu 2016 recommends a higher threshold to weed out spurious variables. We have (modest) evidence that a higher threshold is predictive of extended performance. Although we do not observe the true

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

coefficient estimate, we do see which variables improved with more data and which deteriorated. For all 29 variables, 9 variables, that is, 31%, improved their *t*-statistics in the extended sample, 69% deteriorated. For those variables with absolute *t*-statistics in the original sample of 2.0 or more, 36% improved. At a threshold of 2.5, this increases to 39%. At a threshold of 3.0, it creeps up to 40%. At 3.5, with 7 variables left, it is 43%. At 4.0, two of three variables (gip and gpce, but not vrp) improve, and at 4.5, both remaining variables improve.

Our paper then explores the robustness of these variables. This includes extending their sample periods backward (when available); using a single testing specification for all variables (rather than following the authors' testing recipes); looking at out-of-sample predictive performance and model stability; and reporting hypothetical performance in a number of reasonable investment strategies.

Moving to "homologous" specifications (Table 3), in which variables had to predict the log-equity premium nonoverlapping at their native frequencies, 3 more variables lost their good in-sample performance, leaving 10 variables.[1]

Of these 10 variables, 4 variables also showed good out-of-sample (OOS) performance, imposing the Campbell and Thompson 2008 aid that equity-premium predictions should never be negative (which we dub "OOSCT" for out-of-sample-Campbell-Thompson). They are: tchi, the 14 technical indicators in Neely et al. 2014; shtint, the short-stock interest holdings in Rapach, Ringgenberg, and Zhou 2016; accrul, aggregate accruals in Hirshleifer, Hou, and Teoh 2009; and gpce, fourth-quarter growth in personal consumption expenditures in Møller and Rangvid 2015. In addition, 3 more variables from the pre-GW set also performed well: tby, the Treasury-bill rate (Campbell 1987); i/k the investment-capital ratio (Cochrane 1991); and eqis, equity issuing activity (Baker and Wurgler 2000). Finally, if we do not extend variables backward but only forward, three more variables performed better: ogap, the output gap (Cooper and Priestley 2009); tail, the cross-section based tail-risk (Kelly and Jiang 2014); and avgcor, the average correlation of stock returns (Pollet and Wilson 2010).

We note that Campbell and Thompson 2008 articulate reasons a researcher may want to focus on IS prediction in the longest interval, rather than on IS subsets or on both IS and OOSCT prediction, as our paper does. Their logic usually applies when researchers are confident that the model is stable and known by the user in time. They essentially come down to the fact that *under the model, the model is true*, and tests can then be designed to take advantage of this assumption. (This can be considered one manifestation of a strong theoretical prior.) In such cases, stability or OOSCT prediction tests have lower power than one full IS test. Our own priors—and perhaps those of many, but not all, of our

[1] A number of variables lose significance in homologous regressions because we use nonoverlapping returns at native frequencies rather than overlapping returns with aggregated (lower) frequencies. The other changes typically seem more minor. Moreover, there were 13 variables if we count the 6-month delayed-release versions, too.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

readers—are not as confident. Martin and Nagel 2022 similarly recommend OOS tests, explaining how nonlinear dynamics originating even from Bayesian investors can create temporary spurious predictive in-sample associations.

Other researchers may disagree with some of our interpretations. For example, different researchers may consider IS and OOS performance differently. Is the latter merely a "robustness check" or an "inseparable requirement" for trusting a model? What sample periods should one most trust? Is the most recent evidence more interesting? For example, VIX-based predictors have performed extremely well in the most recent 10 years. The reader should feel free to interpret our evidence differently. Our paper still seeks to offer useful evidence regardless of perspective.

Although good OOSCT R^2 performance suggests that an investor could have constructed a trading strategy that deviated an epsilon from Treasuries toward equities when the prediction was good, and vice-versa, investors may also have considered using variables to time the market more aggressively or via more basic strategies. A good analogy is Fama-Macbeth regressions, which make it easy to establish whether a variable has marginal influence. However, researchers then routinely construct portfolios based on investment rules which are then used to test the investment performance in Black-Jensen-Scholes / Fama-French regressions.[2]

We thus entertained four different investment strategies: either tilting a variable-based conditional investment strategy in favor of equity (75% of the [predicted] time in equities, 25% in Treasuries) or not (50% in each); and either investing more when the variable was stronger (based on a predicted Z-value) or keeping it the same ($1 long or short).

Of all our variables, only three performed well in such investment strategies in that they could beat the *all-equity-all-the-time* investment strategy in a statistically significant manner. All three were based on annual variables (accrul, gpce, and gip) and heavily equity-tilted (though these *t*-statistics were weak: 1.31, 1.54, and 1.35, respectively). Another 16% of our attempted investment strategies could outperform *all-equity-all-the-time* (though with even lower statistical significance).[3] The remaining 84% underperformed *all-equity-all-the-time*. About one-third of these in turn not only underperformed *all-equity-all-the-time* but also lost money even in absolute terms, thus doing worse than simple nonparticipation. An investor would have done better negating the investment strategy signal, going long when the variable suggested going short, and vice versa.

---

[2] Investors could likely implement more cautious strategies that would outperform *all-equity-all-the-time* based on OOSCT performance, if they employed only very marginal changes in their investment allocations. We consider more natural investment strategies that seek to profit from the predictions in an economically meaningful manner, akin to sorted portfolios. This necessarily means that OOSCT performance does not map one-to-one into investment performance.

[3] It is of course well known that stock returns have high variances and it is difficult even for a good investment signal to predict well in the limited time frame available to researchers. This is an intrinsic problem of all equity-premium forecasting.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

It is difficult for a reader to keep track of so many papers, variables, and diagnostics. Thus, following our basic tables, our summary Table 6 provides a quick visual display of many of our findings. Of course, this sacrifices detail and nuance.

The best predictors on many metrics were accrul and gpce. However, accrul's performance derives exclusively from one episode, the dot-com rise and crash around the turn of the millennium. Since then, its variation has been modest, in turn not predicting much either way. gpce was more consistent. (Some monthly variables, such as ogap, performed well if the sample starts after World War II.)

A combination of all predictors, a "smart consensus" in which variables were weighted according to their relative *t*-statistics in time, had good OOSCT performance, too. However, a graph shows that this good performance occurred from the mid-1940s to the mid-1970s. Its performance has been largely nondescript since then. We could also not identify a simple investment strategy based on this consensus forecast that outperformed *all-equity-all-the-time*.

Finally, we briefly look at the certainty equivalence value (CEV) of a risk-averse quadratic investor with a gamma of 5 (also suggested by Campbell and Thompson 2008). No variable could reliably beat (in a statistically significant manner) the *all-equity-all-the-time* strategy or a more risk-averse strategy of holding only 20% equity. However, some variables, especially crdstd and shtint, performed well in economic terms and even beat a strategy that used the historical unconditional mean stock market rate of return as its expectation of future expected rates of return. With 46 variables, this is perhaps not surprising.

1. Performance Criteria

To identify papers, we began with various keyword searches of Google Scholar. We then expanded the original culled candidate list by following their citation references. To be included, a paper had to have at least 100 cites. In total, we reviewed over a thousand candidate papers.

The papers themselves had to have explored time-series OLS forecasting of (excess or raw) returns, either absolute or in logs, and for a broad U.S. stock market index (typically the S&P 500 or a CRSP U.S. index). The authors' forecast horizon had to have been at least one month in advance (i.e., no daily forecasts). We also omitted a number of theoretical and methodological papers and papers forecasting cross-sectional returns. We had to have access to the variables, first to confirm the authors' results within their sample periods and then to be able to extend the variable to 2021. This means we had to exclude variables that were proprietary and not available to us.[4]

[4] In Table 2, we give the original paper more "benefit of the doubt" by trying to follow its methods somewhat more closely than we do in subsequent tables. In two cases involving the volatility-based fear index (BH vrp and BTZ vrp), we had to use updated variables that the authors posted on their websites instead of reconstructing and confirming the authors' variables ourselves.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

This procedure ended up with 24 papers published in prominent finance-related journals (*Journal of Finance* [2], *Journal of Financial Economics* [11], *Journal of Financial Markets* [1], *Journal of Monetary Economics* [2], *Management Science* [1], *Review of Finance* [1], *Review of Financial Studies* [6]). Two more papers (Bekaert and Hoerova 2014, *Journal of Econometrics*, and Martin 2017, *Quarterly Journal of Economics*) were suggested later by readers of earlier drafts. We omitted papers with predictors which were redundant with papers selected here. We also cover variables that were published before Goyal and Welch 2008 only briefly. Most importantly, we did not select papers based on our priors about their performance. Our complete list of papers with their new variables is in Table 1. In the end, we had selected 26 papers with 29 variables from 9 journals.

Our paper diagnostics ask only some intuitive and simple questions:

1. Could we replicate the published IS performance? Could we do so with a simple bivariate in-sample predictive regression? Would the IS performance hold when we extend the sample period forward to 2021? Would it hold if we further extended the sample period backward if earlier data are also available?

   We note that if even the IS coefficient is nowadays insignificant, further consideration today seems unwarranted to us.
2. How stable is the model? That is, does a variable show a sign change in predicting the equity premium within its first vs. second sample halves, either in the author's sample or in our extended sample(s)?
3. Does the model predict equity premiums prevailing in time; that is, does it have positive rolling OOSCT R^2 (as in Goyal and Welch 2008 and subject to the additional restriction of not predicting a negative equity premium as in Campbell and Thompson 2008)? This requires the model to offer basic improvement of the conditional forecasting errors over the unconditional ones (the latter from a simple prevailing equity-premium average model).
4. Would some simple timing investment strategies based on the variable have earned higher rates of return than an *all-equity-all-the-time* unconditional investment strategy? We considered four strategies, based on simple untilted and equity-tilted investment strategies; and based on variable-scaled and unscaled investment amounts.

We do offer statistics on whether we can reject the observed *joint* IS and OOSCT performance under the null hypothesis. However, we do not consider that simultaneous imposition of more than these two criteria (e.g., stability in halves and performance in investment strategies) would recommend *lower* thresholds in the IS and OOSCT requested performance (if one wanted to maintain the same conventional statistical significance levels). That is, we do not view positive answers to most of our questions as necessary

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

requirements. Instead, we view them as useful information and diagnostics, allowing readers to make up their own minds. (They do not have to follow our assessments.)

Furthermore, we are also interested descriptively in two more aspects:

5. Was the variable's good performance limited to some time periods?
6. Would the variable have offered positive ex post utility improvement for a quadratic-utility investor with risk aversion parameter 5, as suggested by Campbell and Thompson 2008?

We did not consider some further concerns:

1. We ignore the fact that some authors using high-frequency data variables had the choice to select their frequencies (say, monthly, quarterly, annually).[5]
2. We ignore the fact that, collectively, the profession has examined many more variables and that the variables we observe are themselves already highly selected (Lo and MacKinlay 1990; Harvey, Liu, and Zhu 2016).
3. We do not know how to judge the predictability of specific financial market episodes. For example, was the COVID bull market (resulting in superior forecasts for some variables, but not others) predictable or not? Empiricists only have access to one realized data series, and, of course, pre-COVID unusual episodes (like the 1973–1975 oil bear markets) also helped establish many variables in the first place.

In sum, a sufficiently skeptical researcher may want to impose even more stringent criteria, especially in light of our recycling of the authors' original data. Equivalently, a sufficiently optimistic researcher with stronger priors on a predictive model may want to discount our evidence and follow the advice of Campbell and Thompson 2008. Although even the Campbell-Thompson perspective would presumably also suggest more skepticism if the IS performance by early 2022 had turned markedly worse.

2. Variables and Tables

2.1 Variables

Table 1 contains the glossary of recently published papers and variables that we investigate. It briefly explains their meaning and sample availability. This will be followed, in more detail, with the paper-by-paper discussion below; and in even greater detail in Internet Appendix Table IA.II.

[5] Simulations suggest that one should use not the 10% significance level of 1.65 when allowing consideration of monthly, quarterly, and annual frequencies, but more appropriately a 10% level of 2.1 on the best of the three.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 1
Glossary of recently published papers and variables

1	Atanasov, Møller, Priestley (JF 2021), *» Consumption Fluctuations and Expected Returns*	.
	pce	aggregate consumption to its trend (1953:q1 – 2020:q4)
2	Bakshi, Panayotov, Skoulakis (JFE 2011), *» Improving the predictability of real economic activity and asset returns with forwa*	.
	impvar	forward implied variances (1996:01 – 2021:12)
3	Bekaert, Hoerova (JE 2021), *» The VIX, the variance premium and stock market volatility*	.
	vp	The VIX squared minus the implied volatility. See also BTZ. (1990:01 – 2010:09)
4	Belo and Yu (JME 2013), *» Household & government investment and the stock market*	.
	govik	public-sector investment (1947:q1 – 2021:q4)
5	Bollerslev, Tauchen, Zhou (RFS 2009), *» Expected Stock Returns and Variance Risk Premia*	.
	vrp	variance risk premium (1990:01 – 2021:12)
6	Chen, Eaton, Paye (JFE 2018), *» Micro(structure) before macro? The predictive power of aggregate illiquidity for st*	.
	lzrt	9 illiquidity measures (1926:01 – 2021:12)
7	Colacito, Ghysels, Meng, Siwasarit (RFS 2016), *» Skewness in Expected Macro Fundamentals and the Predictability of Equity Return*	.
	skew	skewness of GDP growth forecasts (1951:q2 – 2019:q2)
8	Chava, Gallmeyer, Park (JME 2015), *» Credit conditions and stock return predictability*	.
	crdstd	loan officer credit standards (1990:q2 – 2021:q4)
9	Cooper and Priestley (RFS 2009), *» Time-Varying Risk Premiums and the Output Gap*	.
	ogap	output gap of industrial production (1926:01 – 2021:12)
10	Driesprong, Jacobsen, Maat (JFE 2008), *» Striking oil: Another puzzle?*	.
	wtexas	oil price changes (1926:01 – 2021:12)
11	Hirshleifer, Hou, Teoh (JFE 2008), *» Accruals, cash flows, and aggregate stock returns*	.
	accrul, cfacc	aggregate accruals and cash flows (1965 – 2021)
12	Huang, Jiang, Tu, Zhou (RFS 2015), *» Investor Sentiment Aligned: A Powerful Predictor of Stock Returns*	.
	sntm	optimized investor sentiment index (1965:07 – 2018:12)
13	Jones and Tuzel (RFS 2013), *» New Orders and Asset Prices*	.
	ndrbl	new orders to shipments of durable goods (1958:02 – 2021:12)

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 1
Continued

14	Jondeau, Zhang, Zhu (JFE 2019), *» Average Skewness Matters*	.
	skvw	average stock skewness (1926:07 – 2021:12)
15	Kelly and Jiang (RFS 2014), *» Tail Risk and Asset Prices*	.
	tail	tail risk from cross-section (1926:07 – 2021:12)
16	Kelly and Pruitt (JF 2013), *» Market Expectations in the Cross-Section of Present Values*	.
	fbm	single factor from B/M cross-section (1926:06 – 2021:12)
17	Li and Yu (JFE 2012), *» Investor attention, psychological anchors, and stock return predictability*	.
	dtoy,dtoat	nearness to Dow 52-week high (1926:01 – 2021:12)
18	Maio (RF 2013), *» The Fed Model and the Predictability of Stock Returns*	.
	ygap	stock-bond yield gap (1953:04 – 2021:12)
19	Maio (JFM 2016), *» Cross-sectional return dispersion and the equity premium*	.
	rdsp	stock-return dispersion (1926:09 – 2021:12)
20	Mrtn (QJE 2017), *» Expected Return on the market*	.
	rsvix	scaled risk-neutral vix (1996:01 – 2021:12)
21	Møller and Rangvid (JFE 2015), *» End-of-the-year economic growth and time-varying expected returns*	.
	gpce, gip	year-end economic growth (1947/26 – 2021)
22	Neely, Rapach, Tu, Zhou (MS 2014), *» Forecasting the Equity Risk Premium: The Role of Technical Indicators*	.
	tchi	14 technical indicators (1951:02 – 2021:12)
23	Piazzesi, Schneider, Tuzel (JFE 2007), *» Housing, consumption, and asset pricing.*	.
	house	share of housing in consumption (1929 – 2021)
24	Pollett and Wilson (JFE 2010), *» Average correlation and stock market returns*	.
	avgcor	average correlation of daily stock returns (1926:03 – 2021:12)
25	Rapach, Ringgenberg, Zhou (JFE 2016), *» Short interest and aggregate stock returns*	.
	shtint	short stock interest (1973:01 – 2021:12)
26	Yu (JFE 2011), *» Disagreement and return predictability of stock portfolios*	.
	disag	analyst forecast disagreements (1981:12 – 2021:12)

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Our 29 variables can be broadly grouped into six categories:

1. **Macroeconomic:** CP ogap and JT ndrbl (monthly); AMP pce, BY govik, and CGP crdstd (quarterly); CGMS skew (semiannual); and MR gpce (and gip) and PST house (annual).
2. **Sentiment:** HJTZ sntm, Maio ygap, RRZ shtint (monthly); and HHT accrul (and cfacc, annual).
3. **Variance-Related:** BH vp, BPS impvar, BTZ vrp, and Mrtn rsvix (all monthly).
4. **Stock Cross-Section:** CEP lzrt, JZZ skvw, KJ tail, KP fbm, Maio rdsp, PW avgcor (all monthly).
5. **Other Stock Market:** NRTZ tchi, LY dtoy (and dtoat), and Yu disag (all monthly);
6. **Commodities:** DJM wtexas (monthly).

Most market-price-derived variables are monthly. Most macroeconomic variables are quarterly or annual. Financial statement and macroeconomic variables raise yet another concern that we cannot fully address: the variables are often not available in time. The former are typically available only in accounting statements released a few months after the period to which they apply; the latter are often not just released late but even revised for years after the fact. We simply use the authors' choices (typically assuming immediate availability), although we also consider a version delayed by 6 months for our annual variables.

In addition, our paper also briefly revisits the performance of 17 variables already investigated in Goyal and Welch 2008: the dividend-price ratio (d/p), the dividend-yield (d/y), the earnings-price ratio (e/p), the dividend-payout ratio (d/e), as in Campbell 1987; stock volatility (svar), as in Guo 2006; book-market (b/m), as in Kothari and Shanken 1997 and Pontiff and Schall 1998; net issuing activity (ntis), as in Boudoukh et al. 2007; equity issuing activity (eqis), as in Baker and Wurgler 2000; the Treasury-bill rate (tby), as in Campbell 1987; the long-term yield (lty), the long-term bond rate of return (ltr), the term-spread (tms), the default yield (dfy), the default rate of return (dfr), as in Fama and French 1989, the inflation rate (infl), as in Fama and Schwert 1977; private investment (i/k), as in Cochrane 1991, and "cay," as in Lettau and Ludvigson 2001. For precise definitions, please refer to Goyal and Welch 2008.

2.2 Tables

To examine the predictive performance of 46 variables while fitting into the space of a standard article, we have to be frugal in our descriptions. This is best accomplished by following a standard template discussing each variable, while referring to a set of common tables. Our visual summary in Table 6 sometimes makes it easier to maintain a big-picture overview.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Our first task is to confirm that we can create variables that match the performance proposed by the original papers. In many cases, the authors have posted or shared their data series, allowing us to confirm their key results using our own replications of their variable calculations.[6]

Table 2 shows our ability to replicate the basic results of the original paper in the original sample period. In this table, we follow many of the authors' choices. We indicate some choices with flags in our table. (We note when this is not sufficient.) Our first choice in this table is to focus on the variable alone. For almost all variables, controls or more complex author choices are not required to replicate their basic inference. The variables typically hold up quite well in simple univariate[7] regressions with the highest data frequency we have at our disposal (though we never predict at a higher frequency than monthly). Table 2's rightmost columns show the performance when we extend the sample forward to 2021 and when we extend it not only forward but also backward.

Table 3 shows the predictive performance in what we call our "homologous specification." Here, variables had to predict the same target in the same way: log equity premiums using CRSP's definition of the S&P 500 including dividends net of the (Ken French posted) Treasury return in nonoverlapping periods at their native frequencies. The in-sample performance is based on a simple predictive standard OLS regression. The Internet Appendix IA.VI shows only a modest improvement if we predict annual equity premiums for monthly variables.

Table 3 also shows the in-time prevailing out-of-sample performance, similar to our main enterprise in Goyal and Welch 2008.[8] Here, we focus on the in-time OOSCT R^2,

$$R^2_{\text{oos}} = 1 - \frac{\sum_t (r_t - \hat{r}_{t-1})^2}{\sum_t (r_t - \bar{r}_{t-1})^2},$$

where $\hat{r}_{t-1}$ is the conditional forecast at time $t-1$ and $\bar{r}_{t-1}$ is the prevailing mean at time $t-1$. We star this "OOSCT R^2" using bootstrapped MSE-F statistics of McCracken 2007.[9] The variables are *always* constructed on a

[6] The exceptions were BH vp and BTZ vrp, where we simply used the authors' variables themselves. We could extend CGMS skew and HJTZ sntm to 2020, but not to 2021.

[7] We do not replicate authors' full set of multivariate controls or more complex treatments, but we do follow many other choices of the papers.

[8] We do not predict lower-frequency stock returns with higher-frequency predictors. Thus we need not worry about overlapping observations. In a previous draft, we found that higher-frequency variables generally did not do better predicting lower-frequency equity premiums, either with or without overlap.

[9] For bootstrap information, refer to Internet Appendix IA.III. We use the MSE-F statistic for starring OOSCT R^2, because we are interested in population-level predictive ability (whether a variable has any predictive content). One can test finite-sample predictive ability (whether a variable has useful predictive content given that parameters are estimated). Giacomini and White 2006 study such a question in the context of rolling regressions (where the null hypothesis then, necessarily, depends on window length).

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 2
Near replications of papers' predictive results

			Vrbl	Author				Reported		Similar		→	IS X Forw		+ Back
	Pub	Paper	Name	Sample			Codes	Coef b	(T)	Coef b	(T)		Coef b	(T)	=↓
1	2014	BH*	vp*	M	1990:01	2010:09	QO5LX	0.43	2.43^{**}	0.46	2.55^{***}	↘	0.39	2.10^{**}	n/a
2	2011	BPS*	impvar*	M	1998:09	2008:09	MN5LX	9.92	$>3.00^{***}$	8.26	3.60^{***}	↘	4.80	2.02^{**}	1996– = ok
3	2009	BTZ	vrp	M	1990:01	2007:12	QO5LX	0.47	2.86^{***}	0.12	4.35^{***}	↘	0.03	0.72 ✗	n/a ✗
4	2018	CEP	lzrt	M	1948:01	2015:12	MNCLX	2.59	2.85^{***}	2.30	2.51^{***}	↘	0.16	0.38 ✗	1926– = ✗
5	2009	CP	ogap	M	1948:01	2005:12	MNCLX	−0.11	-4.08^{***}	−0.09	-3.67^{***}	↗	−0.10	-4.51^{***}	1926- ✗
6	2008	DJM	wtexas	M	1973:10	2004:04	MNCLX	−0.09	-3.57^{***}	−0.10	-3.19^{***}	↘	−0.03	−1.31 ✗	1926– = ✗
7	2014	HJTZ	sntm*	M	1965:07	2010:12	MN5LX	0.58	3.04^{***}	0.48	2.59^{***}	↗	0.45	2.66^{***}	n/a
8	2013	JT	ndrbl	M	1958:02	2009:12	QOCLX	−0.46	-3.21^{***}	−0.31	-2.47^{**}	↗	−0.29	-2.75^{***}	n/a
9	2019	JZZ	skvw	M	1963:08	2016:12	MNCUX	−0.13	-3.10^{***}	−0.11	-2.58^{***}	↘	−0.07	−1.34 ✗	1926– = ✗
10	2014	KJ	tail	M	1963:01	2010:12	MNCUX	4.54	2.08^{**}	4.97	2.32^{**}	↘	4.15	2.16^{**}	1926- ✗
11	2013	KP	fbm	M	1930:01	2010:12	ANCLR	—	2.85^{***}	0.18	3.33^{***}	↘	0.15	3.12^{***}	1926– = ok
12	2012	LY*	dtoy*	M	1958:01	2009:12	QOCUX	0.32	2.09^{**}	0.22	1.92^{*}	↘	0.12	1.20 ✗	1926– = ✗
13	2012	LY*	dtoat*	M	1958:01	2009:12	QOCUX	−0.48	-3.79^{***}	−0.29	-3.54^{***}	↘	−0.21	-3.17^{***}	1926– = ok
14	2013	$\text{Maio}_{(13)}$	ygap	M	1953:04	2008:12	MNCLX	0.20	2.94^{***}	0.01	1.84^{*}	↘	0.00	0.50 ✗	n/a ✗
15	2016	$\text{Maio}_{(16)}$	rdsp	M	1963:07	2013:12	AOCLX	−3.45	-2.55^{***}	−2.38	-2.30^{**}	↘	−1.70	−1.32 ✗	1926– = ✗
16	2017	Mrtn	rsvix	M	1996:01	2012:12	SO5UX	2.10	2.46^{**}	2.12	2.19^{**}	↗	2.00	2.31^{**}	n/a
17	2014	NRTZ	tchi	M	1951:01	2011:12	MNCLX	0.12	2.12^{**}	0.26	1.88^{*}	↘	0.20	1.52^{*}	n/a
18	2010	PW	avgcor	M	1963:01	2007:01	MNCLX	0.06	2.65^{***}	0.05	2.58^{***}	↘	0.04	2.31^{**}	1926- ✗
19	2016	RRZ	shtint	M	1973:01	2014:12	MN5LX	−0.50	-2.50^{**}	−0.43	-2.15^{**}	↘	−0.37	-1.88^{*}	n/a
20	2011	Y	disag	M	1981:12	2005:12	AOCLX	−0.17	-2.59^{***}	−0.16	-3.98^{***}	↘	−0.03	−1.14 ✗	n/a ✗

(continues)

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 2
Contiuned

			Vrbl	Author				Reported		Similar		→	IS X Forw		+ Back
	Pub	Paper	Name	Sample			Codes	Coef b	(T)	Coef b	(T)		Coef b	(T)	=↓
21	2020	AMP	pce	Q	1953:Q4	2017:Q4	QN5LX	−0.43	-3.28^{***}	−0.45	-3.35^{***}	↗	−0.47	-3.70^{***}	n/a
22	2013	BY	govik	Q	1947:Q1	2010:Q4	QNCLX	1.02	2.11^{**}	1.07	2.06^{**}	↘	0.68	1.46 ✗	n/a ✗
23	2015	CGP	crdstd	Q	1990:Q2	2013:Q4	QNCLX	−0.10	-2.45^{**}	−0.10	-2.20^{**}	↘	−0.07	−1.47 ✗	n/a ✗
24	2016	CGMS	skew	S	1951:S2	2010:S2	SN5LX	−0.02	-2.66^{***}	−0.02	−1.48 ✗	↘	−0.02	−1.17 ✗	n/a ✗
25	2008	HHT	accrul	A	1965	2005	ANCLX	0.07	3.33^{***}	0.06	2.75^{***}	↘	0.05	2.40^{**}	n/a
26	2008	HHT	cfacc	A	1965	2005	ANCLX	−0.05	-2.42^{**}	−0.06	-2.87^{***}	↗	−0.06	-3.14^{***}	n/a
								HHT cfacc Jul-Jun:		−0.05	−1.65 ✗		−0.04	−1.47 ✗	
27	2015	MR	gpce	A	1948	2009	ANCUX	−14.61	-4.88^{***}	−14.91	-4.54^{***}	↗	−15.14	-5.13^{***}	1947– = ok
28	2015	MR	gip	A	1948	2009	ANCUX	−3.78	-5.74^{***}	−3.80	-5.13^{***}	↗	−3.77	-5.26^{***}	1926- ✗
								MR gip Jul-Jun:		0.28	−0.78 ✗		0.45	0.36 ✗	
29	2007	PST	house	A	1936	2001	5O5LX	8.44	3.65^{***}	4.91	2.64^{***}	↗	5.31	2.97^{***}	1929- ✗

† To obtain statistical significance like the authors in their own samples, we had to deviate in some cases from simple bivariate regressions on the latest vintage numbers: The BH vp variable in the replication is the original vintage. (Their updated vintage in the original sample period has an in-sample replication of 0.30 with a *t*-statistic of 1.32.) We kept the sign of sntm, in line with HJTZ's series, rather than negate it. We included control variables for BPS impvar. And we ran LY dtoy and LY dtoat in a trivariate regression.

Explanations: Papers and variables are defined in Table 1. The "author sample" shows the data availability frequency and original sample period. The "reported" statistics appeared in the original paper. The "Similar" is our replication of the original paper's main finding in the same sample period with the same frequency and dependent variable (see codes), but often without control variables and other author refinements. (Unlike here, in later regressions, for comparisons, we switch all forecasting regression to predict the log equity premium and highest available frequencies. The appendix shows the original alternatives.) Most specifications are based only on bivariate regressions with the key independent variable as the only regressor. The "X Forw" extends the sample forward. It is preceded by an uparrow if the *t*-statistic improves in the authors' favor. The "+Back" shows that ogap, tail, avgcor, gip, and house could be extended backward but then lose their inference. Eight other backward-extensible variables retain their inference ("= ok"). Fifteen variables had no further backdata, leaving the final column blank.

Codes: (1) M=Monthly, Q=Quarterly, S=Semi-Annual, A=Annual. (2) N=Non-overlapping, O=Overlapping. (3) C=CRSP, 5=S&P. (4) L=Logged, U=Unlogged. (5) X=Excess, R=Raw Return (KP only).

Interpretation: Simple bivariate regressions with just the predictor are capable of mimicking the authors' key predictive result in all cases (except skew, which could not be replicated, either). Only 9 of 29 variables improved their *t*-statistic as more data (to 2021) came in. Extending the data forward to 2021 removes the IS statistical significance of 12 of 29 variables. Extending it backward, too, removes the IS significance of a further five variables. Two extra rows show that using Jul-to-Jun instead of calendar years fails cfacc and gip annual predictions.

LY, HHT, and MR propose two variables each, leaving 26 papers. We also consider 17 variables from pre-GW papers. Total = 46 variables.

Downloaded from https://academic.oup.com/rfs/article/37/11/34907749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 3
Homologous specification, IS, and OOSCT

Panel A: Monthly variables

Paper	Vrbl Name	In-sample (IS) Coef *b*	NW T	Regression R^2 IS	OOSCT	p(IS,OOSCT)
BH	vp	0.51	1.44	1.47 ✗	−0.23 ✗	
BPS	impvar	0.13	0.35	0.09 ✗	−3.19 ✗	
BTZ	vrp	0.10	0.17	0.06 ✗	0.21	
CEP	lzrt	0.04	0.20	0.01 ✗	0.02	
CP	ogap 1926-	−0.22	−0.66	0.17 ✗	−0.20 ✗	
	1948-	−0.58	−4.42	1.93^{***}	0.19^{*}	✔$0\%^{***}$
DJM	wtexas	−0.27	−1.41	0.26 ✗	0.14^{*}	$6\%^{*}$
HJTZ	sntm	n/c	n/c	1.04^{*}	−0.94 ✗	$7\%^{*}$
JT	ndrbl	−0.27	−1.40	0.41 ✗	−0.47 ✗	
JZZ	skvw	−0.02	−0.07	0.00 ✗	−0.47 ✗	
KJ	tail 1926-	0.04	0.26	0.01 ✗	−0.35 ✗	
	1963-	0.37	2.50	0.75^{*}	0.68^{**}	✔$1\%^{***}$
KP	fbm	n/c	n/c	3.16 ✗	−1.50 ✗	
LY	dtoy	0.17	0.41	0.10 ✗	−0.36 ✗	
LY	dtoat	−0.07	−0.27	0.02 ✗	−0.08 ✗	
$\text{Maio}_{(13)}$	ygap	0.11	0.54	0.06 ✗	−0.77 ✗	
$\text{Maio}_{(16)}$	rdsp	0.29	0.78	0.30 ✗	−0.99 ✗	
Mrtn	rsvix	0.15	0.33	0.12 ✗	−3.27 ✗	
NRTZ	tchi	0.30	1.69	0.51^{*}	0.48^{**}	✔$2\%^{**}$
PW	avgcor 1926-	0.19	0.82	0.13 ✗	−0.18 ✗	
	1963-	0.41	2.15	0.91^{**}	0.19^{*}	✔$3\%^{***}$
RRZ	shtint	−0.37	−1.79	0.67^{*}	1.26^{**}	✔$0\%^{***}$
Y	disag	0.06	0.33	0.02 ✗	−0.20 ✗	
BMRR	ntis	−0.36	−1.44	0.44 ✗	−0.47 ✗	
Cmpl	tby	−0.29	−1.82	0.29^{*}	0.37^{**}	✔$2\%^{**}$
CSa	d/p	0.17	0.71	0.10 ✗	−0.06 ✗	
CSb	d/y	0.22	0.90	0.17 ✗	−0.06 ✗	
CSc	e/p	0.27	1.59	0.26 ✗	−0.64 ✗	
CSd	d/e	−0.11	−0.38	0.04 ✗	−0.93 ✗	
CSe	svar	−0.08	−0.19	0.02 ✗	−0.01 ✗	
FFa	lty	−0.24	−1.57	0.19 ✗	0.25^{**}	✔$3\%^{**}$
FFb	ltr	0.22	1.42	0.17 ✗	−0.82 ✗	
FFc	tms	0.18	1.10	0.11 ✗	0.02	
FFd	dfy	0.06	0.12	0.01 ✗	−0.12 ✗	
FFe	dfr	0.23	0.91	0.19 ✗	−0.30 ✗	
FS	infl	−0.21	−0.95	0.15 ✗	0.10^{*}	$9\%^{*}$
KS	b/m	0.29	0.88	0.28 ✗	−1.06 ✗	

Panel B: Quarterly variables

Paper	Vrbl Name	In-sample (IS) Coef *b*	NW T	Regression R^2 IS	OOSCT	p(IS,OOSCT)
AMP	pce	−1.73	−3.70	4.61^{***}	−0.38 ✗	✔$0\%^{***}$
BY	govik	0.60	1.52	0.59 ✗	0.07	$8\%^{*}$
CGP	crdstd	−1.72	−1.67	4.56 ✗	4.54^{**}	✔$1\%^{**}$
Crn	i/k	−1.65	−3.73	4.41^{***}	3.47^{***}	✔$0\%^{***}$
LL	cay	−0.09	−0.18	0.01 ✗	−6.11 ✗	

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 3
Continued

Panel C: Calendar-year

Paper	Vrbl Name	In-sample (IS) Coef *b*	In-sample (IS) NW T	Regression R^2 IS	Regression R^2 OOSCT	p(IS,OOSCT)
CGMS	skew	0.38	0.16	0.05 ✗	−2.29 ✗	
HHT	accrul	4.76	2.29	8.25**	4.40*	✔3%**
HHT	cfacc	−5.53	−3.09	11.19**	16.35***	✔0%***
MR	gpce	−5.78	−3.61	12.48***	5.43**	✔0%***
MR	gip	−0.28	−0.09	0.02 ✗	−2.05 ✗	
PST	house	1.86	0.96	0.92 ✗	0.46*	9%*
BW	eqis	−5.54	−2.73**	8.17**	1.49*	✔1%**

Panel D: Mid-year (i.e., 6-month delay)

Paper	Vrbl Name	In-sample (IS) Coef *b*	In-sample (IS) NW T	Regression R^2 IS	Regression R^2 OOSCT	p(IS,OOSCT)
CGMS	skew	1.78	0.81	1.37 ✗	−0.43 ✗	
HHT	accrul	5.12	2.61	10.76**	6.31**	✔1%**
HHT	cfacc	−3.29	−1.44	4.51 ✗	12.08**	✔1%**
MR	gpce	−4.25	−1.94	7.50*	8.33**	✔1%**
MR	gip	−0.52	−0.17	0.05*	−0.47 ✗	
PST	house	1.38	0.65	0.36 ✗	0.10 ✗	
BW	eqis	−4.68	−1.76	4.05 ✗	−6.99 ✗	10%*

Explanations: IS Coefficients and IS R^2 are starred based on their bootstrapped values. One star represents 5%–10% significance level. The OOSCT *t*-statistics are starred based on bootstrapped significances of the MSE-F statistic of Clark and McCracken 2001. ✗ represents "not significant" even at the 10% level (including authors' original samples). OOSCT is the Campbell-Thompson zero-lower-bound (but not slope-sign restricted) OOS R^2. p(IS,OOSCT) is a conjoint significance under the null hypothesis (shown when better than 10%). We checkmark values better than 5%, one-sided (i.e., equivalent to a *t*-statistic of about 1.645 under the normal distribution). "n/c" for HJTZ and KP represents "not comparable," because the authors used PLS rather than OLS.

Interpretation: *Panel A:* vp, impvar, ogap, ndrbl, fbm, dtoat, and rsvix lose IS significance in this synchronized log-equity premium specification, even though they had forward and backward in-sample significance in Table 2 (where we tried to stay closer to author specifications). Variables tchi, shtint, and d/p, continue to perform well in-sample. (sntm and fbm are in-sample optimized.) The standout IS/OOSCT performers were, in order of predictive IS and OOSCT performance: shtint, tby, tchi, and d/p. wtexas, sntm, and infl had decent performance, too. *Panel B:* Most quarterly predictors (pce,i/k, crdstd, and possibly govik) performed well. cay did not. *Panels C+D:* accrul, cfacc, and gpce performed well even with reporting delay. Without reporting delay, eqis performed well, too.

real-time basis—for example, when variables require filters or regression coefficients for construction (such as pce), the required calculations are always based only on prevailing historical values. As already mentioned, our focus is only on the OOSCT variant of the out-of-sample R^2, in which we impose the restriction that the equity-premium prediction should not be negative. This has good theoretical justifications and pragmatically should help in our sample because this sample experienced high equity premiums.

Our OOS period always starts 20 years after the IS period, but never earlier than 1946. Authors can reasonably object that there are good reasons they started their own analyses earlier or later. Obviously, different starting periods

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

can lead to different results, just like different ending periods. Our own choices were the same as those in Goyal and Welch 2008 and largely dictated by our desire to keep the same scheme across our 29+17 variables. Importantly, our figures make it easy to assess how a different starting period would affect the results.

Table 3 also offers a "p(IS,OOSCT)" conjoint statistic that shows how often we would expect to see both IS and OOSCT performance as high as what we observe in the sample under the null hypothesis of no association. Adding the further requirement of OOSCT performance on top of IS performance suggests using smaller critical *t*-statistics. For example, a variable that has both an IS p-level of 6% and an OOSCT p-level of 6% may have a conjoint p-level under 5%. (The two probabilities cannot be multiplied, because they are not independent.)

In **Appendix Table A.1**, we show the coefficients when the samples were divided into two halves. Compared to rolling OOSCT R^2, the "IS halves" coefficients use all sample periods. This makes them more powerful diagnostics of model stability, although investors in time would be more interested in OOSCT performance. Our visual summary Table 6 includes a rather mild form of stability diagnostics based on observed sign changes, too.

As shown in the Internet Appendix IA.VI, the results are very similar if we predict simple rather than log equity premiums.[10] We experimented with more sophisticated forecasting, but the inference was similar enough to recommend the brevity and simplicity of an exposition based on plain OLS forecasting techniques. This includes our consideration of forecasting and techniques from Kostakis, Magdalinos, and Stamatogiannis 2015 and Cederburg, Johnson, and O'Doherty 2023.[11]

In **Table 4**, we summarize how well a risk-neutral investor would have performed in seeking to time her investments based on these variables. (The table only shows better performers. Complete results are in Internet Appendix Tables IA.1-IA.4.) The benchmark unconditional investment strategy is earning the equity premium itself. We name this investment strategy *all-equity-all-the-time*. The investment strategy performance is always based on zero-investment strategies (i.e., either the stock market financed with bills or the opposite,

[10] For monthly variables, vrp, wtexas, and dtoat improved in OOCTCT R^2, just crossing a one-sided 10% significance level. Only avgcor improved significantly. For annual variables, house improved in OOSCT R^2 from 0.46% (insignificant at 90%) to 0.56% (significant at 90%) in simple returns. Its June prediction OOSCT R^2 turned negative.

[11] We do however recommend both. The latter further looks at a good number of recent prediction variables. Recent finance literature investigates the pitfalls associated with multiple-hypothesis testing. The common approaches are to control family-wise error rate (Romano and Wolf 2005; White 2001) or false discovery rate (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). **However, these approaches are not suitable for the nested models that we study here.** We thank Todd Clark and Michael McCracken for clarifying these issues for us.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 4
Investment performance better than *all-equity-all-the-time*

Panel A: Untilted $1-Unscaled Investment Strategy

	Vrbl		Conditional R(V)			#Obs		Unconditional R (U)			Δ (= V – U)	
	Ppr	Name	Long	Short	L–S	Bull	Bear	L(Eq)	S(TB)	L – S	Mean	SR
J	MR	gpce	12.0	4.6	7.5	32	22	12.0	4.6	7.4	0.1	0.01 ✔

Panel B: Untilted Z-Scaled Investment Strategy

	Vrbl		Conditional R(V)			#Obs		Unconditional R (U)			Δ (= V – U)	
	Ppr	Name	Long	Short	L–S	Bull	Bear	L(Eq)	S(TB)	L – S	Mean	SR
M	RRZ	shtint	16.7	1.5	15.2	237	111	16.1	2.2	13.9	1.3	0.06 ✔
M	Y	disag	15.3	1.0	14.4	227	14	15.2	1.1	14.2	0.2	0.07 ✔
M	FFb	ltr	12.1	4.3	7.8	446	466	11.7	4.7	7.0	0.8	0.03 ✔
M	FS	infl	3.5	2.0	1.5	413	499	3.4	2.1	1.3	0.2	0.01 ✔
Q	CGP	crdstd	4.5	−1.2	5.7	48	39	2.1	1.3	0.8	5.0	0.11 ✔
D	HHT	accrul	3.8	0.3	3.5	15	22	2.4	1.7	0.7	2.8	0.09 ✔
J	HHT	accrul	3.1	0.0	3.1	15	21	1.6	1.6	0.1	3.1	0.11 ✔
D	HHT	cfacc	7.0	0.6	6.4	25	12	5.8	1.8	4.1	2.3	0.18 ✔
D	MR	gpce	5.9	1.9	4.0	32	23	5.7	2.1	3.6	0.4	0.03 ✔
J	MR	gpce	6.7	1.1	5.6	32	22	5.7	2.1	3.7	1.9	0.15 ✔

Panel C: Long-Equity-tilted $1-Scaled Investment Strategy

	Vrbl		Conditional R(V)			#Obs		Unconditional R (U)			Δ (= V – U)	
	Ppr	Name	Long	Short	L–S	Bull	Bear	L(Eq)	S(TB)	L – S	Mean	SR
M	NRTZ	tchi	12.0	4.2	7.8	481	131	11.9	4.3	7.6	0.3	0.01 ✔
M	Y	disag	10.7	0.8	10.0	240	1	10.3	1.2	9.2	0.8	0.22 ✔
Q	AMP	pce	12.7	4.0	8.8	179	14	12.4	4.4	8.0	0.8	0.08 ✔
Q	CGP	crdstd	9.9	0.4	9.5	68	19	8.8	1.5	7.4	2.1	0.10 ✔
D	HHT	cfacc	14.7	1.9	12.8	32	5	13.5	3.2	10.3	2.5	0.18 ✔
J	HHT	cfacc	13.2	2.7	10.6	31	5	12.7	3.1	9.6	1.0	0.08 ✔
D	BW	eqis	13.6	3.4	10.3	64	11	13.0	3.9	9.1	1.2	0.10 ✔
J	MR	gpce	12.4	4.2	8.2	46	8	12.0	4.6	7.4	0.8	0.06 ✔

namely, bills financed by shorting value-weight stocks).[12] We are interested in investment strategies that are timed conditional on the variable. When the timing investor is bullish (i.e., in the market), the unconditional and conditional strategies invest and earn the same. When the timing investor is bearish, the conditional strategy earns the opposite of the unconditional strategy.

Considering specific investment strategy variants is a new aspect of our paper. We do so based on *scaled* and *unscaled* timing strategies, and *equity-tilted* and *untilted* timing strategies:

1. The *untilted, unscaled* timed investment strategy invests either $1 in the market when it is bullish (financed with Treasuries) or $1 in

[12] Zero-investment strategies can always be viewed as "overlays." Thus, they are comparable but never mutually exclusive.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 4
Continued

Panel D: Long-Equity-tilted Z-Scaled Investment Strategy

	Vrbl		Conditional R(V)			#Obs		Unconditional R (U)			Δ (= V – U)	
	Ppr	Name	Long	Short	L–S	Bull	Bear	L(Eq)	S(TB)	L – S	Mean	SR
M	NRTZ	tchi	16.1	5.2	10.9	481	131	15.7	5.5	10.2	0.7	0.04 ✔
M	RRZ	shtint	22.8	1.8	21.1	286	62	21.9	2.7	19.1	2.0	0.13 ✔
M	Y	disag	20.7	1.5	19.2	240	1	20.7	1.5	19.2	0.1	0.22 ✔
M	Cmpl	tby	8.9	3.9	5.0	446	466	8.6	4.2	4.3	0.7	0.02 ✔
M	FFb	ltr	14.9	3.9	11.0	594	318	13.9	4.9	9.0	2.0	0.09 ✔
M	FFc	tms	12.9	3.9	8.9	612	300	12.5	4.3	8.3	0.7	0.05 ✔
M	FS	infl	7.3	1.2	6.1	708	204	5.5	1.9	4.6	1.5	0.10 ✔
Q	CGP	crdstd	11.6	-0.1	11.7	68	19	10.0	1.5	8.5	3.3	0.11 ✔
D	HHT	accrul	8.8	0.1	8.7	31	6	6.5	2.4	4.1	4.6	0.20 ✔
J	HHT	accrul	7.8	-0.2	8.0	31	5	5.3	2.2	3.1	4.9	0.22* ✔
D	HHT	cfacc	12.2	1.8	10.4	32	5	11.3	2.8	8.5	2.0	0.19 ✔
D	MR	gpce	9.8	2.2	7.7	46	9	9.5	2.6	6.9	0.8	0.12 ✔
J	MR	gpce	11.1	1.8	9.3	46	8	10.3	2.5	7.8	1.4	0.21* ✔
D	MR	gip	8.5	1.9	6.6	74	2	8.5	1.9	6.6	0.0	0.16* ✔
D	BW	eqis	11.2	3.1	8.0	64	11	11.0	3.3	7.8	0.3	0.05 ✔

Explanations: Variables and sample periods are defined in Table 1. Investment begins 20 years after the sample has started. "M" is monthly; "Q" is quarterly; "D" is calendar year; and "J" is mid-year investing (i.e., with reporting delay). All measures are annualized, incl. the Sharpe Ratio (SR). The starring of OOSCT performance follows Lo 2002. This table shows only performances no worse than the unconditional strategy. The appendix contains the complete set of tables with results.

Panel A: The conditional timing strategy invests $1 in the equity premium financed by the T-bill if the predictive full-sample coefficient is positive and V is above its median, or the coefficient is negative and V is below its median (both "bullish"); and the opposite otherwise ("bearish"). The unconditional strategy is always bullish and earns the equity premium. *Panel B:* Here, the investment is not $1, but $Z (V realization minus prevailing median, divided by prevailing standard deviation), that is, the signal strength influences not just the direction but also the amount of the investment. The unconditional strategy is always bullish and invests |$Z|. *Panel C:* the decision cutoff is not the median but the 25th percentile (a more bearish cutoff). Thus, the strategy typically invests more optimistically in equities. (Please note that Y's positive performance is based on only one month, in which it would have been bearish.) *Panel D:* The combination of B and C.

Interpretation: Of 53×4 original rows representing different investment strategies—53 include December and June specifications and original GW variables, in Tables IA.1-IA.4)—about about 16% (34) suggest investment performance that would have exceeded the buy-and-hold performance. Three would have done so in a statistically significant manner: accrul (T of 1.31) and gpce (T of 1.54) with an investment delay of 6 months, using a long-equity tilted and Z-scaled strategy, and gip (T of 1.35) using a long-equity tilted and Z-scaled strategy. skew and crdstd have economically good mean performance and Sharpe ratios.

Treasuries financed by shorting the stock market when it is bearish. This conditional strategy decides based on whether the variable is bullish or bearish by looking above or below its historical median in time, according to the sign of the full-sample coefficient.

2. The *equity-tilted* strategies switch from long stocks to long Treasuries only if the signal is much more bearish, that is, at the 25th percentile rather than the median.
3. The *scaled* strategies first calculate a Z-like score in time; that is, they subtract the prevailing *median* (when untilted) and 25th percentile (when equity-tilted) and divide by the prevailing standard

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

deviation.[13] This determines the investment amount. For example, if the prevailing forecasting coefficient is positive (so being above the x cutoff [median or first quartile] means bullish), if the Z-score calculates –0.5, the conditional strategy would short \$0.50 in the market and purchase \$0.50 of Treasury bills. The comparative unconditional strategy would long \$0.50 in the market and purchase \$0.50 in Treasury bills.

4. The equity-tilted scaled strategy is a combination of the two preceding choices.

The descriptions of individual papers in Section 3. refer to the above tables.

We offer three more tables, already noted in the introduction. **Table 5** considers the CEV of a risk-averse investor. **Table 6** provides a visual overview of many of our results. Many readers will find this "forest perspective" a better starting point that the more detailed "tree perspective." And **Table 7** shows the performance of at attempt to combine variables, based on their to-date performance (Zarnowitz and Lambros 1987; Rapach, Strauss, and Zhou 2010).

3. Empirical Performance

This section runs through each paper's performance. It describes the performance of the variables proposed in each recent paper, sorted by data frequency and alphabetical name of first author. Our standard discussion template for papers presents each variable as follows:

1. A modified version of the original abstract that focuses on relevant aspects. For the complete version, please refer to the original paper.
2. A basic intuitive explanation of the variable and sample period. (This explanation is commonly insufficient to replicate our version of the variable. The fully detailed discussion appears in the original papers and in our Internet Appendix.)
3. A discussion of the performance in four parts: **[A]** IS performance, using the author's specification. **[B]** The homologous specification performance IS; **[C]** The same homologous specification but for the OOSCT R^2; **[D]** the performance in our four simple investment strategies; and **[E]** a discussion of a time-series graph of the prediction performance. (The graphs' out-of-sample performance also imposes the zero equity-premium lower bound.) Importantly, the figures do not

[13] Using the median instead of the mean in the Z-score assures that the Z-score is aligned with being bullish or bearish. For a calculation example, say that the distribution of a positive predictor is from –1 to +1, the 25th percentile is –0.5; the median is 0; the 75th percentile is +0.5; and the standard deviation is 1.0. If the variable is at its 40th percentile equal to –0.2, in the untilted strategy, the signal is bearish (below the median). The Z-score would be $(-0.2-0)/1=-0.2$. This means that we will short \$0.2 in the stock market and long \$0.2 in the risk-free rate. For the tilted strategy, the signal would be bullish (above the 25th percentile). The Z-score would be $(-0.2-(-0.5))/1=0.3$. We would long \$0.3 in the market and short \$0.3 in bonds.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

graph R^2 but sum-squared errors. Thus, a variable has horizontal rather than downward-drifting performance when it loses forecasting power. Specifically, we plot

$$y_t = \sum_{s=1}^{t} \left[(r_s - \bar{r}_{s-1})^2 - (r_s - \hat{r}_{s-1})^2 \right],$$

where $\hat{r}$ is the prevailing model prediction and $\bar{r}$ is the prevailing historical equity premium average. Note that these are not divided by t, and thus there is no natural tendency for y to trend toward zero if the model predicts as well as the prevailing mean. A model that is predictive should thus have a y_t that is trending up. As already mentioned in the introduction, OOSCT and investment performance do not line up perfectly. Real-world strategies are based on rules that map signals into investment strategies, analogous to the way that Fama-Macbeth based signals are mapped into (often sorted) portfolios, which are then examined in a Black-Jensen-Scholes / Fama-French time-series analysis.

4. Our somewhat subjective assessment ("evaluation"). Whereas our own priors are skeptical, readers with stronger priors in the models can come to different conclusions.

3.1 Monthly Variables

3.1.1 BH: Bekaert and Hoerova 2014 As our first paper, this discussion is modestly more extensive than those for later ones. We begin with a summary statement quoting from the paper, followed by the variable description, followed by its performance summary.

Abstract: *Using more plausible estimates of the variance premium and stock market volatility, ... the well-known results in BTZ exaggerate the predictive power of the variance premium for stock returns. However, the equity variance risk premium remains a reliable predictor of stock returns.*

Variable: Unlike many other variables, we did not compute vp ourselves but relied on author-posted data. vp is obtained by subtracting from the prevailing VIX-squared a measure of previously prevailing fitted volatility based on such variables as earlier realized variance and VIX^2 measures. Over time, the authors have been changing the variable retrospectively, too. Thus, the original-vintage data still produces a statistically significant coefficient in the original sample, but the current-vintage data no longer does. The current-vintage data however produces statistically significant coefficients in data ending in 2021. As is common in many papers, the BH paper also entertains other extended variants of its variable. However, the simplest version of the BH variable in itself performs *better* than what the authors reported, with a T-statistic of 2.55 (rather than 2.43).

Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of vp in the original sample period (–2010), with a

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

T of 2.55. Extended forward to 2021 (and now using their latest vintage data), vp retains its significance but weakens[14] moderately (T of 2.10). The lack of a cross-out mark in the last column means that extending vp backward does not affect the inference (here positive statistical significance). In this case, the authors provided only one extra year of data that we could add. Having confirmed that we can replicate the authors' IS inference and that it persists if we extend the data forward to 2021, we can continue.

[B] In our homologous specification (–2021) in Table 3, the IS coefficient is no longer statistically significant, with an IS T of 1.44.[15]

Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (We investigate only the homologous specification for OOSCT performance. In later descriptions, we parenthesize the description of OOSCT and investment performance when the homologous IS performance is already insignificant.)

Table A.1, Panel A shows that the T-statistic halved in the latter half of the sample period. This is not a sign-change, so it is not particularly concerning. However, the temporal decline helps explains the lower (non-significant) IS T in the extended sample.

[C] Table 3 shows that the OOSCT R^2 of vp is negative. There is no entry in the final column, because we are never listing conjoint significance levels (p(IS,OOSCT)) that are above 10%.

[D] The investment performance of vp was poor. In these cases, Table 4 does not describe any performance details. An interested reader needs to refer to the Internet Appendix tables.

Internet Appendix Table IA.1 shows that in our first strategy — where one would have invested $1 in equities financed by the risk-free rate whenever the variable predicted an equity premium above its historical median and vice-versa — the vp-based strategy returned 8.0%/year less than *all-equity-all-the-time*. More remarkably, this strategy actively selected periods for equity before stock prices fell: it even lost money in absolute terms (–0.9%/year). Reversing the vp signal for investment purposes would have been better.

The other three investment strategies underperformed *all-equity-all-the-time*, but did not lose money in absolute terms. Table IA.2 shows that if one had scaled the investment by the Z-score of the vp-model, vp would have returned 3.2% year less than *all-equity-all-the-time*. Table IA.3 shows that vp would have returned 3.4% less than *all-equity-all-the-time*, investing $1 whenever the variable was above its 25th percentile (thus tilting the conditional strategy more in favor of equity). Finally, Table IA.4 shows that vp would have returned

[14] In general, weakening should not be overread, but T-statistics are expected to increase as more data is sampled. This would of course not always be so.

[15] The difference in T between the 2.10 in Table 2 and the 1.44 in Table 3 is mostly due to the authors' use of overlapping quarterly data instead of monthly data. Using non-overlapping monthly returns but keeping the rest of their specification drops the T to 1.58. The remaining 0.14 is due our use of log returns, etc.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

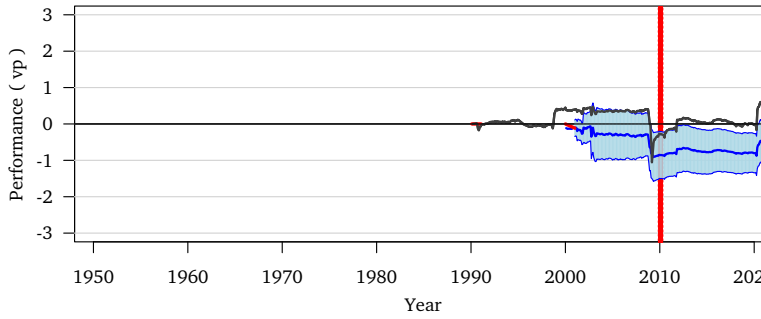

Figure 1
IS and OOS Predictive Performance of BH vp (monthly)
Explanations: The black line is IS performance; the blue line is OOS performance (with ±2 standard errors drawn based on Diebold-Mariano, which usually corresponds well to the more appropriate Clark-McCracken MSE-t statistic in the table). Unlike the tables, these figures are based on simple rather than log returns. (When this makes a difference, it is noted in the text.) Note that the lines do not graph R^2 but sum-squared cumulative errors. (This means that a horizontal continuation does not imply consistent prediction (as it would if this graphed R^2), but no further marginal improvement. In this example, vp predicted better OOS than the mean model from 2020 to 2022, but neither outpredicted nor underpredicted it from 2012 to 2019.)

0.8% less than *all-equity-all-the-time*, using both the Z-score and equity-tilt modifications.

[E] Our performance figures (as in Goyal and Welch 2008) show when a variable performed well and when it did not. Intuitively, in these figures, when the prediction based on the conditioning variable (here vp) does well, the line increases; when the variable underperforms (the prevailing mean for the OOSCT lines), the line decreases. The black line is the IS predictions, the blue line is the OOSCT prediction (which means the conditional prediction in time is compared to the unconditional prediction at time t, the prevailing mean). A variable that is statistically significant should lie solidly above the zero horizontal line.[16] The authors' original end of sample is shown with a vertical red line.

Figure 1 shows that vp predicted poorly in the Great Recession bear market. It predicted well in the dot-com and the Covid bull markets. The performance recovered well during Covid but not enough to render the overall OOSCT performance positive. This is also partly the case because the required incept delay meant that the OOSCT performance just missed the good performance in 2000.

Evaluation: We dismiss vp as a useful predictor of equity premiums, primarily based on lack of IS performance in our homologous specification; and secondarily, poor OOSCT performance. It also would have offered poor investment performance. Presumably, if the evidence in Bekaert and Hoerova

[16] The blue range is the ±2 standard deviation range for OOSCT prediction, based on an MSE-T statistic Diebold and Mariano 1995, which is related to but not identical to the MSE-F statistic used to star the OOSCT R^2 in the tables.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

2014 was consistent with "the equity variance risk premium remaining a reliable predictor," the extended evidence should now be viewed as somewhat less consistent. We note again that all of our "evaluations" are more subjective, allowing the reader to disagree with our own assessment.

3.1.2 BPS: Bakshi, Panayotov, and Skoulakis 2011 Abstract: *[BPS] present an option positioning that allows [them] to infer forward variances from option portfolios. The forward variances [they] construct from equity index options help to predict ... (iii) stock market returns... .*
Variable: BPS synthesize the exponential of integrated variance using a strip of European calls and puts, written on the market index.
Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of impvar in the original sample period (–2008), with a T of 3.60. Extended forward to 2021, impvar weakens (T of 2.02). The variable could be extended backward but only for two years. The inference does not change, which is indicated by the '= ok' in the final column. We note that, unlike for most other variables, to retain IS statistical significance, it was necessary to include the authors' controls in Table 2. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with a T of 0.35. Its poor performance in this specification is not unexpected, due to the just-mentioned need for the additional author controls in the replication. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of impvar is negative. **[D]** The investment performance of impvar was poor. In the two strategies not tilted towards equity, impvar not only does not beat *all-equity-all-the-time*, it even loses money in absolute terms. When tilted towards equity and unscaled, it barely manages to avoid such exceptionally bad performance.) **[E]** Figure 2 shows a second reason why our results are so different from the authors': impvar collapsed completely in the Great Recession, just after the BPS sample had ended in Sep 2008. Specifically,

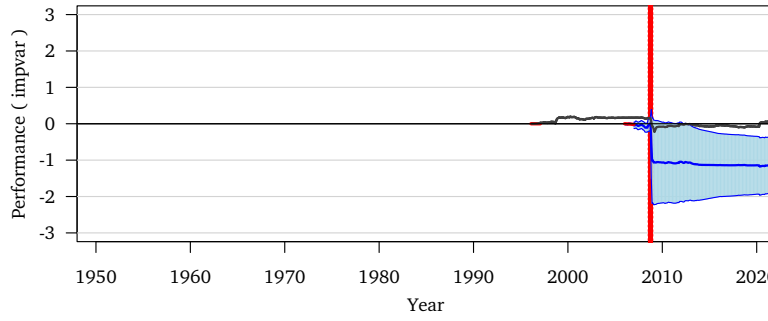

Figure 2
IS and OOS Predictive Performance of BPS impvar **(monthly)**

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

impvar's Sep and Oct 2008 values failed to predict the –18% and –8.5% drops in the value-weighted market rate of return in Oct and Nov 2008.
Evaluation: We dismiss impvar as a useful predictor of equity premiums, based on its poor IS performance (and, further, poor OOSCT performance). Presumably, if the evidence in Bakshi, Panayotov, and Skoulakis 2011 was consistent with "a role for implied volatility," the extended evidence should now be viewed as somewhat less consistent.

3.1.3 BTZ: Bollerslev, Tauchen, and Zhou 2009 Abstract: *Motivated by the implications from a stylized self-contained general equilibrium model incorporating the effects of time-varying economic uncertainty, [BTZ] show that the difference between implied and realized variation, or the variance risk premium, is able to explain a nontrivial fraction of the time-series variation in post-1990 aggregate stock market returns, with high (low) premia predicting high (low) future returns. [The] empirical results depend crucially on the use of "model-free," as opposed to Black-Scholes, options implied volatilities, along with accurate realized variation measures constructed from high-frequency intraday as opposed to daily data. The magnitude of the predictability is particularly strong at the intermediate quarterly return horizon...*

BTZ is the most-cited paper in our set, with about 1,500 Google scholar citations. Bekaert and Hoerova 2014 offers a further discussion of BTZ and improvement of the vrp measure.
Variable: Unlike many other variables (but like BH's vp), we did not recompute vrp ourselves. Instead, we used the vrp data updated regularly by the authors themselves and posted on their website.
Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of vrp in the original sample period (–2007). However, the IS extended forward to 2021 drops to a T of only 0.72. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant. The model is unstable. The coefficient turns from positive to negative in our second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of vrp is positive, but small and not statistically significant. **[D]** The investment performance of vrp was poor. It was between 2.6%/year and 7.0%/year worse than *all-equity-all-the-time*.) **[E]** Figure 3 shows that vrp did well following the Great Recession. However, its performance turned poor in early 2020. (Of course, as in many other cases, the authors could not have known because this occurred after they had published their paper. Not shown, the pure OOS performance without the zero truncation did *far* worse, dropping off the chart.)
Evaluation: We dismiss vrp as a useful predictor of equity premiums, primarily based on poor IS performance, secondarily based on poor OOS performance. The investment performance was poor, too. Presumably, if the evidence in Bollerslev, Tauchen, and Zhou 2009 was consistent with their "stylized

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

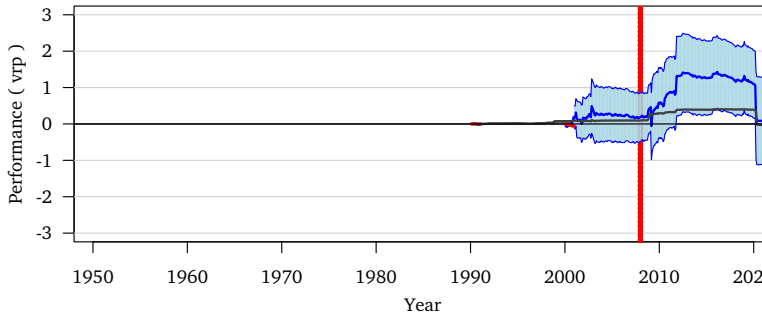

Figure 3
IS and OOS Predictive Performance of BTZ vrp **(monthly)**

self-contained general-equilibrium model with time-varying economic uncertainty based on the same analysis," the extended evidence should now be viewed as somewhat less consistent.

3.1.4 CEP: Chen, Eaton, and Paye 2018 Abstract: *[CEP] constructs and analyzes various measures of trading costs in US equity markets covering the period 1926-2015. These measures contain statistically and economically significant predictive signals for stock market returns and real economic activity. [They]...find strong evidence that the component of illiquidity uncorrelated with volatility forecasts stock market returns...*
Variable: lzrt is the log of the number of zero returns. The series has structural break adjustments for tick-size reductions in 1997 and 2001 (these are included by regressing the series on binary variables that take the value of 1 after the tick-size reductions, and 0 otherwise, then taking the residuals as the final series).
Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of lzrt in the original sample period (–2015). However, the IS extended forward to 2021 drops to a T of only 0.38. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant. This IS falls to a T of 0.20. The model is unstable. The coefficient turns from positive to negative in our second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of lzrt is positive, but small and not statistically significant. **[D]** The investment performance of lzrt was poor. Without a heavy equity tilt, lzrt even loses money in absolute terms. With equity tilt, lzrt still greatly underperforms *all-equity-all-the-time*.) **[E]** Figures 4 illuminates the performance. Chen, Eaton, and Paye 2018 caught lzrt nearly at its best. Nasdaq came online in 1973, increasing the number of missing daily returns and thereby caught the 1974 bear market. lzrt also outperformed in the Great Recession. However, lzrt missed the subsequent bull market as well as the Covid bull market. Otherwise, lzrt was fairly unremarkable.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

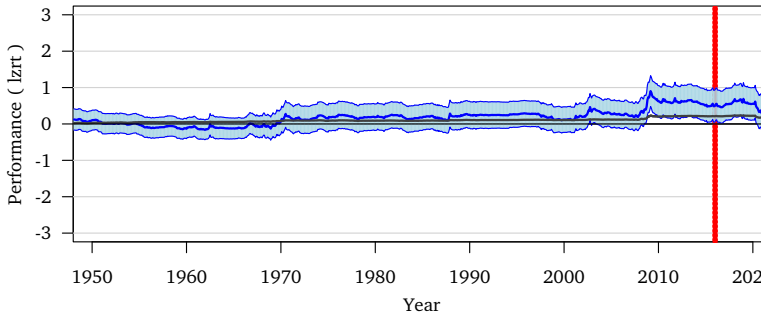

Figure 4
IS and OOS Predictive Performance of CEP lzrt (monthly)

Evaluation: We dismiss lzrt as a useful predictor of equity premiums, based on poor IS performance. Presumably, if the evidence in Chen, Eaton, and Paye 2018 was consistent with "a role for illiquidity," the extended evidence should now be viewed as somewhat less consistent.

3.1.5 CP: Cooper and Priestley 2009 Abstract: *[CP show that] the output gap, a production-based macroeconomic variable, is a strong predictor of U.S. stock returns. It is a prime business cycle indicator that does not include the level of market prices, thus removing any suspicion that returns are forecastable due to a "fad" in prices being washed away. The output gap forecasts returns both in-sample and out-of-sample, and it is robust to a host of checks...*
Variable: The output gap (ogap) is the deviation of the log of industrial production from a trend that incorporates both a linear and a quadratic term.
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of ogap in the original sample period (–2005), with a T of –3.67. With more data to 2021, the T even strengthens to –4.51. However, once its sample is extended backwards to 1926, ogap no longer predicts well even IS. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant. Including the pre-CP period, the IS T is –0.66. Thus, if extended to the longest sample, further OOSCT investigation would seem unwarranted. However, if we restrict the sample to extend forward only (starting after World War 2), ogap remains a good predictor of performance. (**[C]** Table 3 shows that the OOSCT R^2 of ogap is negative. However, if not extended backward, ogap has a positive OOSCT R^2, modestly significant at the one-sided 8% level. The joint p(IS,OOSCT) statistic is highly significant. **[D]** The investment performance of ogap was poor. It always underperforms *all-equity-all-the-time*. Even if not extended backwards, all four investment strategies underperform *all-equity-all-the-time*.) **[E]** Figure 5 shows that the IS performance was steady. However, the OOSCT performance early on was very poor (i.e., the blue line starts at

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

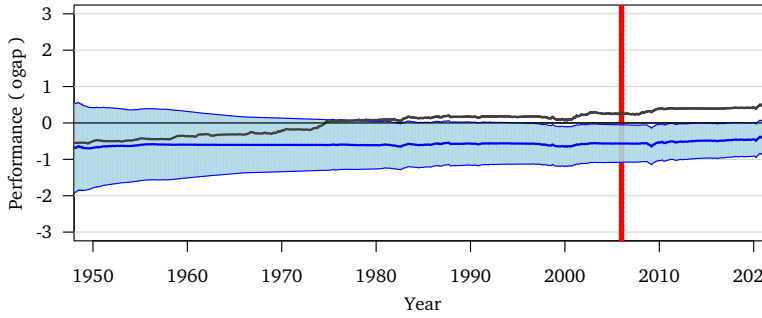

Figure 5
IS and OOS Predictive Performance of CP ogap **(monthly)**

–0.7%), so the (slow, steady, but small) improvements from 1950 to 2021 are insufficient to make much difference. The ogap simply has no longer moved much either way. It usually predicts about the same equity premium as the equity premium's historical average.

Evaluation: We would not outright dismiss ogap as a useful predictor of equity premiums. ogap performed well in IS and OOSCT predictions when only extended forward. It was small and steady. However, it did not offer superior investment strategies. Furthermore, if ogap is extended backwards, too, then it is easy to dismiss.

3.1.6 DJM: Driesprong, Jacobsen, and Maat 2008 Abstract: *[DJM show that] changes in oil prices predict stock market returns worldwide...These results cannot be explained by time-varying risk premia as oil price changes also significantly predict negative excess returns. Investors seem to underreact to information in the price of oil. A rise in oil prices drastically lowers future stock returns. Consistent with the hypothesis of a delayed reaction by investors, the relation between monthly stock returns and lagged monthly oil price changes strengthens once we introduce lags of several trading days between monthly stock returns and lagged monthly oil price changes.*

Variable: wtexas is the price of West-Texas Intermediate crude oil, as obtained from *Global Financial Data* services. We also extend the sample backward from 1973, when Driesprong, Jacobsen, and Maat 2008 begin.

Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of wtexas in the original sample period (–2004). However, the IS T extended forward to 2021 drops to only –1.31. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with a T of –1.41. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of wtexas is positive (0.14, statistically significant at the 6% level). **[D]** The investment performance of wtexas was poor. It underperformed *all-equity-all-the-time*.) **[E]** Figure 6 shows that wtexas had

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

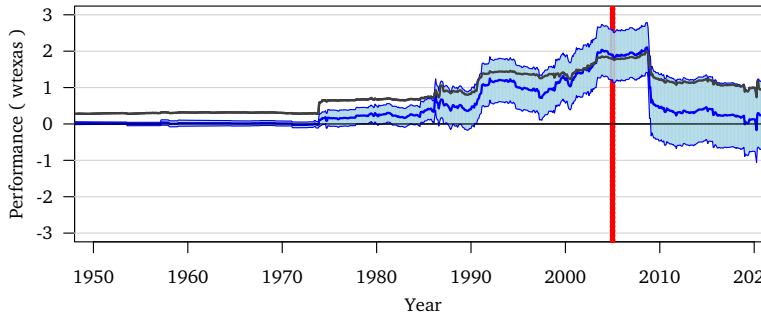

Figure 6
IS and OOS Predictive Performance of DJM wtexas **(monthly)**

good annual OOSCT R^2 performance until the Great Recession. It collapsed in June 2008, when the oil price dropped from \$139/b to \$39/b and has underperformed ever since. The latter occurred just after Driesprong, Jacobsen, and Maat 2008 was published, which explains the difference between their results and our own.

Evaluation: We dismiss wtexas as a useful predictor of equity premiums, based on its poor IS performance. Presumably, if the evidence in Driesprong, Jacobsen, and Maat 2008 was consistent with "models of delayed reaction by investors (offering simple high trading profits)," the extended evidence should now be viewed as somewhat less consistent

3.1.7 HJTZ: Huang et al. 2015 Abstract: *[HJTZ] propose a new investor sentiment index that is aligned with the purpose of predicting the aggregate stock market. By eliminating a common noise component in sentiment proxies, the new index has much greater predictive power than existing sentiment indices have both in and out of sample, and the predictability becomes both statistically and economically significant. In addition, it outperforms well-recognized macroeconomic variables and can also predict cross-sectional stock returns sorted by industry, size, value, and momentum. The driving force of the predictive power appears to stem from investors' biased beliefs about future cash flows.*

HJTZ can be viewed as combining the sentiment measure of Baker and Wurgler 2007, which was designed for the cross-section and not for market timing, with the in-sample optimization method of Kelly and Pruitt 2013.

Variable: sntm uses the Baker and Wurgler 2007 six sentiment variables, but weights them to optimize the predictive performance in sample using the technique pioneered in Kelly and Pruitt 2013.

Figure 7 plots the time-series of sntm. Sentiment was very pessimistic in 1968-1969, 1982, and 2000-2001; and very optimistic in 1974-1976. Oddly, sentiment does not have intuitive time-series behavior. Figure 7 shows that sntm was not particularly optimistic in 1998–1999 (though it did collapse later

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

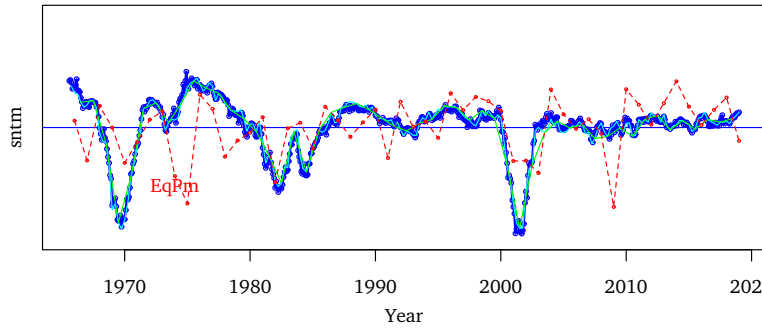

Figure 7
Time-Series of Sentiment (sntm) and Equity Premia

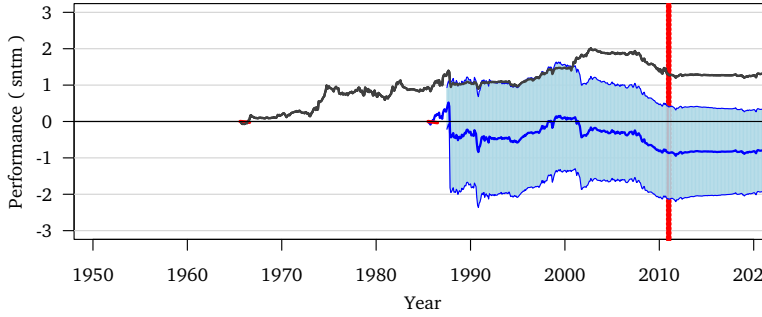

Figure 8
IS and OOS Predictive Performance of HJTZ sntm (monthly)

in 2001–2002), and that it has remained fairly steady throughout the Great Moderation, and the Great Recession. (The data did not reach into Covid.)

Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of sntm in the original sample period (–2010). The T's are 2.59 and 2.66 in our basic and forward-extended samples. However, this is somewhat meaningless because sntm is optimized to maximize the IS performance following Kelly and Pruitt 2013. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is (of course) still statistically significant. **[C]** Table 3 shows that the OOSCT R^2 of sntm is negative. The conjoint significance is 7%, primarily due to the artificially high IS performance using the PLS procedure.[17] **[D]** sntm always underperforms *all-equity-all-the-time*. **[E]** Figure 8 shows that the IS-optimized performance was good from inception to about 2001. Its OOSCT performance, missing the

[17] Although the mean bootstrapped IS T-statistic is also 1.70, the observed T statistic of 2.66 is high enough to remain statistically significant.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

first 20 years of good IS performance, turned poor almost immediately. Both IS and OOSCT performance have been poor at least since 2001.
Evaluation: We dismiss sntm as a useful predictor of equity premiums, primarily because of its difficult-to-assess IS performance coupled with poor OOSCT performance, especially since the end of the dot-com bubble. Presumably, if the evidence in Huang et al. 2015 was consistent with "a new investor sentiment index aligned to predict the stock market," the extended evidence should now be viewed as somewhat less consistent

3.1.8 JT: Jones and Tuzel 2013 Abstract: *[JT] investigate the asset pricing and macroeconomic implications of the ratio of new orders (NO) to shipments (S) of durable goods. NO/S measures investment commitments by firms, and high values of NO/S are associated with a business cycle peak. We find that NO/S proxies for a short-horizon component of risk premia not identified in prior work. Higher levels of NO/S forecast lower excess returns on equities...at horizons from one month to one year. These effects are generally robust to the inclusion of common return predictors and are significant on an out-of-sample basis as well...*
Variable: ndrbl is the ratio of new orders to shipments of durable goods, obtained from the Census Bureau. Jones and Tuzel 2013 interpret their variable as a forecast of future investment growth.
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of ndrbl in the original sample period (–2009). With more data, the T improves modestly from –2.47 to –2.75. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant.[18] **[C]** Table 3 shows that the OOSCT R^2 of ndrbl is negative.[19] **[D]** The investment performance of ndrbl was poor. **[E]** Figure 9 shows that much of the good IS performance was due to the performance in 1974–1975, i.e., predicting the oil-shock bear market. This was too early to be included in our OOSCT prediction. Since then, ndrbl has also been unremarkable even IS. In the aftermath of the Tech collapse, ndrbl performed poorly, mispredicting 2001 and 2002, which turned its OOSCT R^2 negative. Otherwise, with modest spikes in the Great Recession and a good spike in the Covid year, ndrbl was mostly unremarkable.
Evaluation: We dismiss ndrbl as a useful predictor of equity premiums, based on its poor IS performance in our homologous specification (and poor OOSCT

[18] This is primarily due to our specification using monthly non-overlapping returns rather than quarterly overlapping returns.

[19] Our OOS periods always begin 20 years after a variable is available. In contrast, Jones and Tuzel 2013, Table 8 start after 5 or 10 years. Thus, with a data start of 1958, they still include the stellar oil-crisis 1974–1975 performance of ndrbl (see Figure 9), whereas we do not. Looking back to Table 2, the model coefficients also declined over their sample, though they did not do so in a statistically significant manner. JT also use quarterly frequency data for their OOS prediction, whereas we remain with the frequency of the main results, monthly. The reader can thus consider the OOS performance of JT to be sensitive rather than negative.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

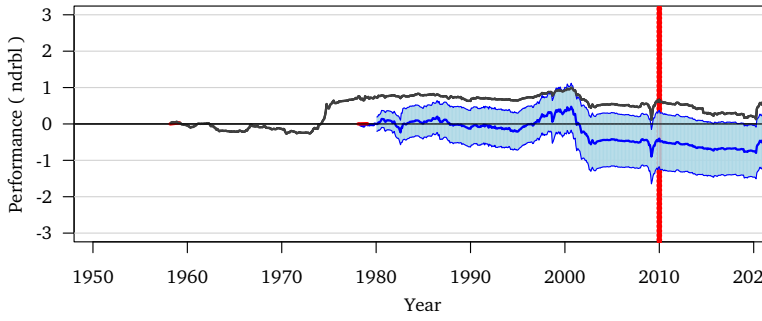

Figure 9
IS and OOS Predictive Performance of JT ndrbl **(monthly)**

performance). Presumably, if the evidence in Jones and Tuzel 2013 was consistent with "the asset pricing implications of new orders to shipments of durable goods," the extended evidence should now be viewed as somewhat less consistent

3.1.9 JZZ: Jondeau, Zhang, and Zhu 2019 Abstract: *[JZZ find that] average skewness, which is defined as the average of monthly skewness values across firms, performs well at predicting future market returns. This result still holds after controlling for the size or liquidity of the firms or for current business cycle conditions. [They] also find that average skewness compares favorably with other economic and financial predictors of subsequent market returns. The asset allocation exercise based on predictive regressions also shows that average skewness generates superior performance.*
Variable: skvw is as described in the authors' abstract. Note that this is *not* the average time-series skewness of the market index itself, but a cross-sectional skewness.
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of skvw in the original sample period (–2016). However, the IS T extended forward to 2021 diminishes to only –1.34. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of –0.07. The model is unstable. The coefficient turns from positive to negative in our second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of skvw is negative. **[D]** The investment performance of skvw was poor. It even lost money in absolute terms in our non-equity-tilted investment strategies.) **[E]** Figure 10 shows that skvw always performed poorly.
Evaluation: We dismiss skvw as a useful predictor of equity premiums, based on its poor IS (and OOSCT) performance. Presumably, if the evidence in Jondeau, Zhang, and Zhu 2019 was consistent with "a role for average of individual skewnesses," the extended evidence should now be viewed as somewhat less consistent

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

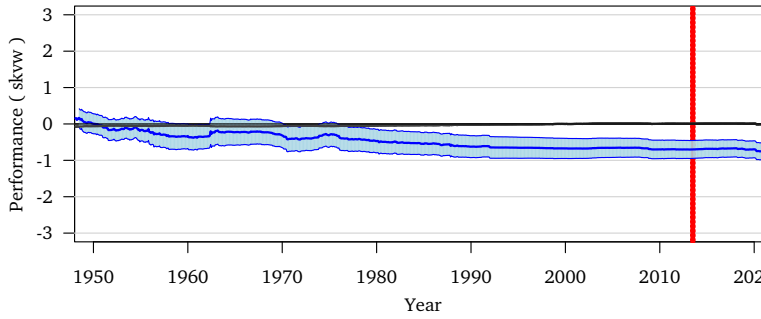

Figure 10
IS and OOS Predictive Performance of JZZ skvw **(monthly)**

3.1.10 KJ: Kelly and Jiang 2014 Abstract: *[KJ] propose a new measure of time-varying tail risk that is directly estimable from the cross-section of returns. [They] exploit firm-level price crashes every month to identify common fluctuations in tail risk among individual stocks. [The] tail measure is significantly correlated with tail risk measures extracted from S&P 500 index options and negatively predicts real economic activity. We show that tail risk has strong predictive power for aggregate market returns.*
Variable: tail is as described in the authors' abstract. Note that tail is *not* the tail risk of the market index itself, but a cross-sectional statistic.
Performance: **[A]** Table 2 can confirm the positive and statistically significant IS coefficient of tail in the original sample period (–2010), with a T of 2.32. Extended forward, tail weakens (T of 2.16). However, once its sample is extended backwards (to 1926 instead of 1963), tail no longer predicts well even IS. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of 0.26. The model is unstable. The coefficient switches from negative to positive in the second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. However, if extended forward only, tail remains a solid statistically significant predictor. (**[C]** Table 3 shows that the OOSCT R^2 of tail is negative. This is again due to our backward extension of the sample. Beginning in 1945, the OOSCT R^2 is strongly statistically significant at the 1% level. **[D]** The investment performance of tail was poor with or without backward sample extension.) **[E]** tail offered no remarkable performance or episodes. The variable barely budged and the predictive coefficient was not large, which is why the performance relative to the equity premium remains rather flat and unremarkable — but steady.
Evaluation: We would not outright dismiss tail as a useful predictor of equity premiums. It had poor IS and OOSCT performance when extended backward. It performed well when only extended forward.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

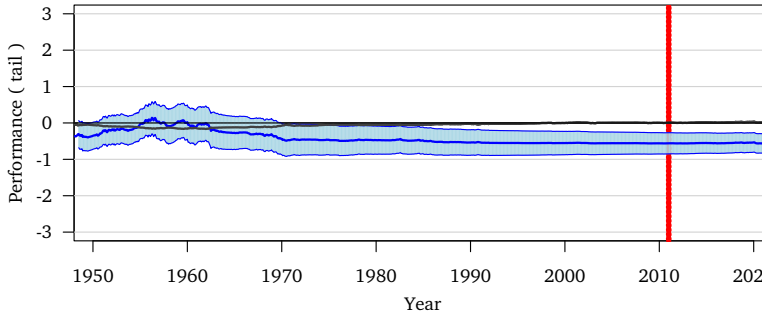

Figure 11
IS and OOS Predictive Performance of KJ tail **(monthly)**

3.1.11 KP: Kelly and Pruitt 2013 Abstract: *[KP find that] returns and cash flow growth for the aggregate U.S. stock market are highly and robustly predictable. Using a single factor extracted from the cross-section of book-to-market ratios, [they] find an out-of-sample return forecasting R^2 of 13% at the annual frequency (0.9% monthly)... [They] present a model linking aggregate market expectations to disaggregated valuation ratios in a latent factor system. Spreads in value portfolios' exposures to economic shocks are key to identifying predictability and are consistent with duration-based theories of the value premium.*

de Oliveira 2022 show that Kelly and Pruitt 2013 is sensitive to a number of implementation choices, although this is in turn disputed by Kelly and Pruitt 2022.

Variable: KP constructs its variable based on a partial least squares (PLS) technique that extracts a latent factor most relevant for predicting returns by exploiting the relationship between the predictors and the returns being forecast. fbm is an outlier in terms of its predictive IS performance, because it was optimized (fitted in PLS) based on ex-post data to do so. Although the bootstrap destroys the natural order of equity premia (by construction), the IS coefficient in each bootstrapped sample still remains optimized to deliver IS predictability in that bootstrapped sample. This causes the bootstrapped IS T-stats to be really high (with a mean of 3.95). The actually observed T of 3.33 is thus not high. This explains the insignificance due to a low bootstrapped p-value (of 0.96) in our Table 3.

Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of fbm in the original sample period (–2010). However, this is somewhat meaningless because fbm is optimized to maximize the IS performance. Not shown, the bootstrap suggests that this high IS performance also occurs frequently under random shuffling. **[B]** As in the authors' original specification, Table 3 shows a high IS R^2 that is common even under the null hypothesis. It is not statistically significant when bootstrapped.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

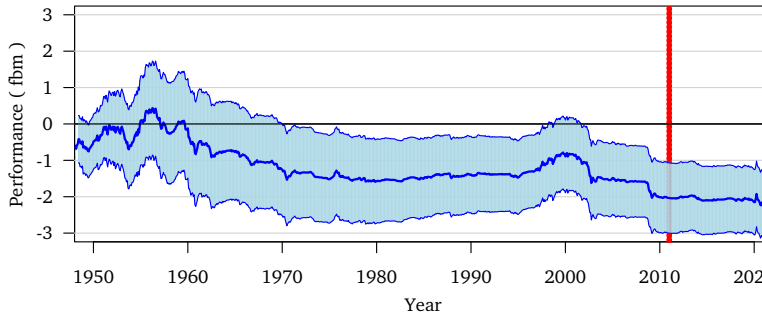

Figure 12
IS and OOS Predictive Performance of KP fbm **(monthly)**

[C] Table 3 shows that the OOSCT R^2 of fbm is negative.[20] **[D]** The investment performance of fbm was poor. **[E]** Figure 12 shows consistently inferior OOSCT performance over the entire sample period. (It makes no sense to plot fbm IS, because it is highly optimized in this respect.) The evidence seems to suggest that fbm is simply overfitted.
Evaluation: We dismiss fbm as a useful predictor of equity premiums, based on its poor IS performance in our homologous specification (and poor OOSCT performance). Presumably, if the evidence in Kelly and Pruitt 2013 was consistent with "high and robust predictive ability and a model linking aggregate market expectations to disaggregated valuation ratios (with spreads in value portfolios being key together with duration-based theories of the value premium)," the extended evidence should now be viewed as somewhat less consistent

3.1.12 LY: Li and Yu 2012 Abstract: *Motivated by psychological evidence on limited investor attention and anchoring, [LY] propose two proxies for the degree to which traders under- and overreact to news, namely, the nearness to the Dow 52-week high and the nearness to the Dow historical high, respectively. [LY] find that nearness to the 52-week high positively predicts future aggregate market returns, while nearness to the historical high negatively predicts future market returns....*
Variable: LY introduce two variables: dtoy and dtoat. The former is the scaled current difference to the 52-week high of the Dow Jones index, the latter is

[20] Further unreported analysis shows that the discrepancy in reported OOS between KP and ourselves can be traced to three issues: [1] their OOS prediction started in 1980, our's in 1946 (for consistency across all papers); [2] their OOS prediction ended in 2010, our's in 2021; and [3] they predict (log) market returns, we predict (log) equity premia. They do so because they try to take advantage of the Cochrane identity. (With its different target, one might argue that KP does not fit into our set of papers being analyzed.) The positive OOSCT R^2 also effectively disappears if we subtract the inflation rate instead of the risk-free rate. Each of these matters and together they explain why they have a positive OOS R^2 and we have a negative one.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

the current difference to the lifetime high. Because the Dow-Jones was mostly moving up, the distance was often near its maximum of 1. The variables are available monthly.

However, the variables are significant even IS only if they are both included simultaneously. They are not significant individually.

Distance to Historical Maximum Price of the Dow-Jones index (dtoy)
Performance: **[A]** Table 2 can confirm the positive and statistically significant IS coefficient of dtoy in the original sample period (–2009). (Again, this requires including both dtoy and dtoat — by themselves, both variables are insignificant.) However, extended forward to 2021, the IS T diminishes to only 1.20. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of 0.41 — not surprising, because we do not include *both* variables. The coefficient is small and turns from positive to negative in the second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of dtoy is negative. **[D]** The investment performance of dtoy was poor. dtoy always underperforms *all-equity-all-the-time*. In the simplest version (untilted, equal investments), it even loses money in absolute terms.) **[E]** Given the poor performance if dtoy is included without dtoat, we do not graph it.
Evaluation: We dismiss dtoy as a useful predictor of equity premiums, based on its poor IS performance (and further poor OOSCT performance). Presumably, if the evidence in Li and Yu 2012 was consistent with "models of psychological evidence on limited investor attention and anchoring," the extended evidence should now be viewed as somewhat less consistent.

Distance to Maximum Price Lifetime (dtoat)
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of dtoat in the original sample period (–2009). Unlike dtoy, the variable remains IS significant when extended (and, of course, included together with dtoy). **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of –0.27. The poor performance in the homologous specification is not surprising, given that the specification required simultaneous inclusion with dtoat, which we do not do in Table 3. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of dtoat is negative. **[D]** The investment performance of dtoat was poor. dtoat not only underperforms *all-equity-all-the-time*, the untilted strategies even lose money in absolute terms.) **[E]** Given the poor performance once dtoat is included without dtoy, we do not graph it.
Evaluation: We dismiss dtoat as a useful predictor of equity premiums, based primarily on poor IS and OOSCT performance (without including both highly correlated variables simultaneously). Presumably, if the evidence in Li and

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Yu 2012 was consistent with "models of psychological evidence on limited investor attention and anchoring," the extended evidence should now be viewed as somewhat less consistent.

Further Assessment

The authors discuss their (highly correlated) variables individually and do not seem to suggest that they require the presence of one another. Our own paper's purpose is not to advance theories by introducing novel predictors (e.g., by considering the difference between dtoy and dtoat). Instead, we strive for consistent treatment across evaluated papers. On our standard single-variable metrics, the two variables fail.

3.1.13 Maio(13): Maio 2013 Abstract: *The focus of this article is on the predictive role of the stock-bond yield gap—the difference between the stock market earnings (dividend) yield and the 10-year Treasury bond yield—also known as the "Fed model". The results show that the yield gap forecasts positive excess market returns...the yield gap has reasonable out-of-sample predictability for the equity premium when the comparison is made against a simple historical average, especially when one imposes a restriction of positive equity premia....An investment strategy based on the forecasting ability of the yield gap produces significant gains in Sharpe ratios.*

Variable: Maio 2013 calculates the Fed model as the dividend-price ratio net of the 10-year government bond yield, the latter multiplied by 10. We use a corrected definition, which removes the multiplication factor.

Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of ygap in the original sample period (–2008). However, the IS T extended forward to 2021 drops to only 0.50.[21] **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of 0.54. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of ygap is negative. **[D]** The investment performance of ygap was poor. The strategies that were not tilted towards equity even lost money in absolute terms.) **[E]** Figure 13 shows that ygap had no remarkable IS performance. The OOSCT prediction performed poorly in the 1970s. Thereafter, it has been largely non-descript, doing worse until about 2000 and better thereafter.

Evaluation: We dismiss ygap as a useful predictor of equity premiums, based on its poor IS performance (and further poor OOSCT performance). Presumably, if the evidence in Maio 2013 was consistent with "the Fed Model

[21] We used the corrected definition — not multiplying the annualized yield by 10 — in Table 2, which reduced the T-statistic from 2.88 to 1.84. Not shown, even if we used the incorrect definition the IS statistic would drop to insignificance when extended to 2021.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

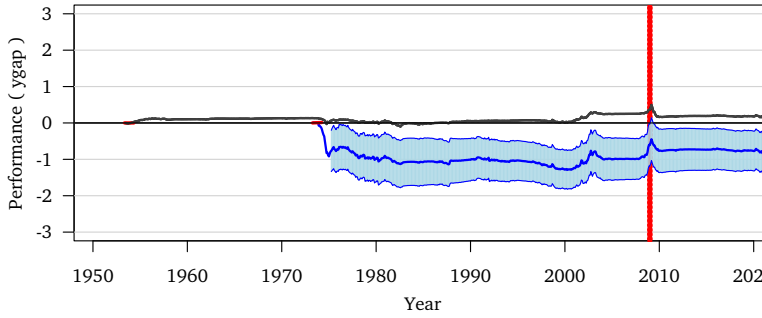

Figure 13
IS and OOS Predictive Performance of Mgap ygap **(monthly)**

as a predictor," the extended evidence should now be viewed as somewhat less consistent.

3.1.14 Maio(16): Maio 2016 Abstract: *[Maio] examines whether stock return dispersion (RD) provides useful information about future stock returns. RD consistently forecasts a decline in the excess market return at multiple horizons, and compares favorably with alternative predictors used in the literature. The out-of-sample performance of RD tends to beat the alternative predictors, and is economically significant as indicated by the certainty equivalent gain associated with a trading investment strategy.*
Variable: rdsp is the cross-sectional standard deviation on the set of 100 size and book-to-market portfolios.
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of rdsp in the original sample period (–2013). However, extended forward to 2021, it loses statistical significance with its T of –1.32. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of +0.78. (The sign switches in non-overlapping regressions.) The model is unstable, too. Its IS coefficient switches sign in our second half. (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of rdsp is negative. **[D]** The investment performance of rdsp was poor. Three strategies lost money in absolute terms. All versions underperformed *all-equity-all-the-time*.) **[E]** Figure 14 shows that rdsp performed poorly IS and OOSCT from 1950 to the mid 1960s, and again from the mid 1980s to 2020. It performed well — but not well enough — in the Covid episode.
Evaluation: We dismiss rdsp as a useful predictor of equity premiums, based on its poor IS performance (and further poor OOSCT performance). Presumably, if the evidence in Maio 2016 was consistent with "a role for useful information in stock dispersion," the extended evidence should now be viewed as somewhat less consistent.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

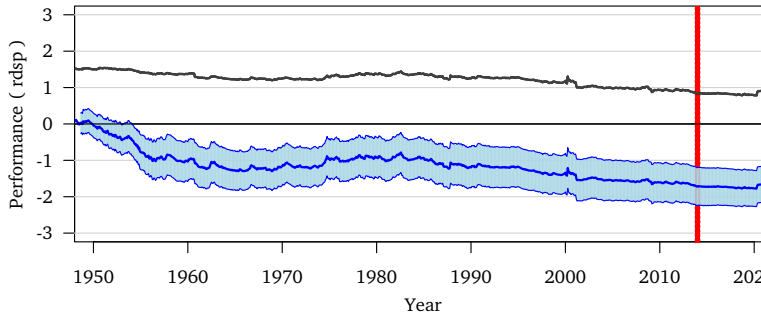

Figure 14
IS and OOS Predictive Performance of Mdsp rdsp **(monthly)**

3.1.15 Mrtn: Martin 2017 Abstract: *[Mrtn] uses the SVIX index as a proxy for the equity premium and argues that the high equity premia available at times of stress largely reflect high expected returns over the very short run.*

The relationship between the squared implied standard deviation and expected returns makes great sense. We note that implied volatilities from equity options had earlier been used to establish bounds on the equity premium in Martin 2011, Backus, Chernov, and Zin 2014, Welch 2016 and Seo and Wachter 2019. Remarkably, Martin 2017 was first to test whether the constraint could be binding, i.e., whether the implied volatility bounds could be related directly to the equity premium.

Variable: Mrtn measures his predictor rsvix as the risk-neutral variance index, i.e., $SVIX^2$. The 1-month rsvix has 99.5% correlation with the more common squared CBOE volatility index (VIX), which is also based mostly on a 1-month horizon. Mechanically, predicting equity premia with the 1-month rsvix index is therefore functionally equivalent to predicting it with the squared VIX. Mrtn uses different horizon rsvix to predict different horizon equity premia: 1-month, 2-month, 3-month, 6-month, and 12-month ahead equity premia. (The longer horizon rsvix series still has 94% correlation with the squared 1-month VIX squared and performs roughly as well in predicting future equity premia as the same-horizon rsvix numbers.) As a volatility index, rsvix is available on a daily basis, but we only consider it on a monthly basis.

Importantly, a closer reading of Martin 2017 shows that the paper itself suggests no statistical significance. In Appendix IA.V, we explain how Martin 2017 interpreted its findings in a different manner. (The difference is primarily about the choice of frequency and about what the null hypothesis is and what the alternative needs to reject.) This appendix also shows that the SVIX variable has performed "too good" in the most recent 10 years, rejecting the original Martin hypothesis with coefficients much above 1.

Performance: **[A]** Table 2 can confirm the positive and statistically significant IS coefficient of rsvix in the original sample period (–2021). Extended forward

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

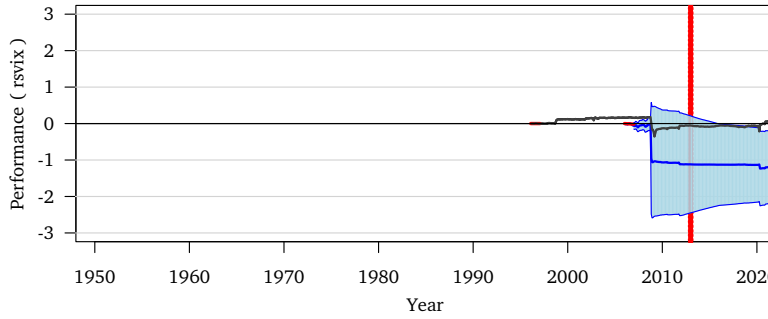

Figure 15
IS and OOS Predictive Performance of Mrtn rsvix **(monthly)**

to 2021, its T improves from 2.19 to 2.31. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with a T of 0.33. The model is unstable. Its IS coefficient switches from negative to positive in our second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of rsvix is negative. **[D]** The investment performance of rsvix was poor. In fact, it performed between 1.1%/year and 10.4%/year(!) worse than *all-equity-all-the-time*.) **[E]** Figure 15 shows that rsvix was modestly successful from inception to 2007 but collapsed badly in the Great Recession. It did not change much thereafter.
Evaluation: We dismiss rsvix as a useful predictor of equity premiums, based on its poor IS performance in our homologous specification (and poor OOSCT performance).

3.1.16 NRTZ: Neely et al. 2014 Abstract: *Technical indicators display statistically and economically significant in-sample and out-of-sample predictive power, matching or exceeding that of macroeconomic variables.*
Variable: tchi is the first principal component of 14 technical indicators, themselves principally versions of moving price averages, momentum, and ("on-balance") dollar-trading volume.
Performance: **[A]** Table 2 can confirm the modest positive and statistically significant IS coefficient of tchi in the original sample period (–2011). However, the IS T extended forward to 2021 drops to only 1.52. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, though its IS T is still a modest 1.69. Thus, it is unclear whether further investigation is warranted. **[C]** Table 3 shows that the OOSCT R^2 of tchi is positive. Moreover, barely missing IS significance and with good OOSCT significance, the joint p(IS,OOSCT) statistic for tchi *is* statistically significant at the 2% level. **[D]** tchi underperformed *all-equity-all-the-time* in untilted $1 strategies and outperformed *all-equity-all-the-time* in equity-tilted strategies (though not in a statistically significant manner). **[E]** Figure 16

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

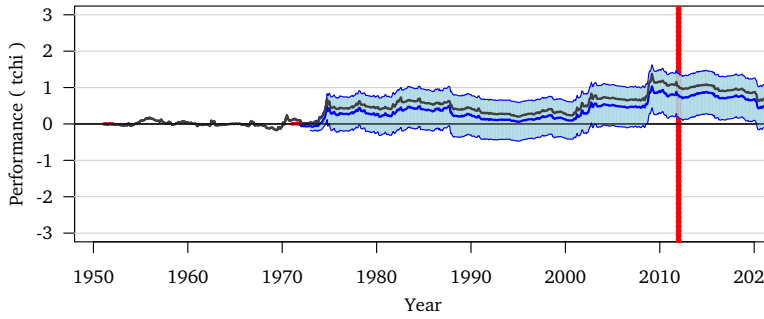

Figure 16
IS and OOS Predictive Performance of NRTZ tchi **(monthly)**

shows that tchi predicted well in late 2008 and early 2009. It reached its brief high point in the Great Recession, i.e., in Feb 2009, predicting Mar 2009. Since 2009, tchi has consistently underperformed. This also explains why our findings differ from those in NRTZ, which ended in 2011.
Evaluation: tchi had marginal IS performance and consistently poor performance since the Great Recession. However, we would think it still deserves actively watching forward.

3.1.17 PW: Pollet and Wilson 2010 Abstract: *...[PW show that] higher aggregate risk can be revealed by higher correlation between stocks. [PW] show that the average correlation between daily stock returns predicts subsequent quarterly stock market excess returns.*
Variable: avgcor is the average correlation among the 500 largest stocks (by capitalization). The daily pairwise correlations of stock returns are multiplied by the product of both stocks' weights relative to total sample market capitalization, then summed to create the measure.
Performance: **[A]** Table 2 can confirm the positive and statistically significant IS coefficient of avgcor in the original sample period (–2007), with an IS T of 2.58. Extended forward, avgcor weakens (T of 2.31). However, once its sample is extended backwards, avgcor no longer predicts well even IS. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of 0.82 (Table 3). However, if extended forward only, avgcor remains statistically significant with a T-statistic of 2.15. **[C]** Table 3 shows that the OOSCT R^2 of avgcor is negative, extended forward and backward. Extended forward only, the OOSCT R^2 has one-sided statistical significance at the 5% level. Together with good IS performance, this is enough to reach statistical significance in p(IS,OOSCT) at the 3% level. **[D]** The investment performance of avgcor was poor. Strategies also performed worse than *all-equity-all-the-time* if *not* extended backwards. **[E]** Figure 17 shows that avgcor had good performance from World War 2 to the mid-2000s. Thereafter, it had its ups and downs. It performed poorly in the Great Recession

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

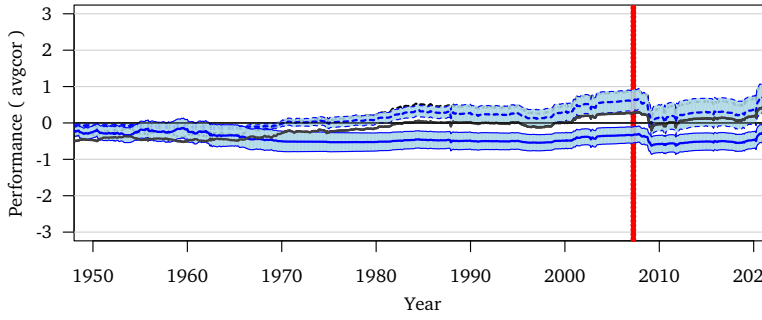

Figure 17
IS and OOS Predictive Performance of PW avgcor **(monthly)**
The dashed lines are simple returns

bear market but recovered in the Covid bull market. However, avgcor is our only variable that actually improves with simple returns. Thus, we also plot the simple return version into the graph.

Evaluation: This is better left to readers. avgcor performed well in IS and OOSCT regression when only extended forward. That is, it performed well from 1945 to the mid-2000s. Thereafter, it no longer did. Even if only extended forward, from 1945 on, avgcor-based investment strategies underperformed *all-equity-all-the-time*. Finally, if avgcor is extended backwards, too, then it is easy to dismiss all around.

3.1.18 RRZ: Rapach, Ringgenberg, and Zhou 2016 Abstract: *[RRZ] show that short interest is arguably the strongest known predictor of aggregate stock returns. It outperforms a host of popular return predictors both in and out of sample, with annual R2 statistics of 12.89% and 13.24%, respectively. In addition, short interest can generate utility gains of over 300 basis points per annum for a mean-variance investor... Overall, our evidence indicates that short sellers are informed traders who are able to anticipate future aggregate cash flows and associated market returns.* (This hypothesis further requires that other intelligent investors ignore publicly available short interest information.)
Variable: shtint is the aggregate short interest in the stock market, calculated as the log of the equal-weighted mean of short interest (as a percentage of shares outstanding) across publicly listed US stocks.
Performance: **[A]** Table 2 can confirm the negative and statistically significant IS coefficient of shtint in the original sample period (–2014), with a T of –2.15. Extended forward, shtint weakens (T of –1.88). **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is still statistically significant, though with a modest T of –1.79. **[C]** Table 3 shows that the OOSCT R^2 of shtint is positive, with our highest monthly R^2 (of 1.26%). Taking both IS and OOSCT significance into account, the p(IS,OOSCT) statistic confidently suggests a predictive association. **[D]** The investment

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

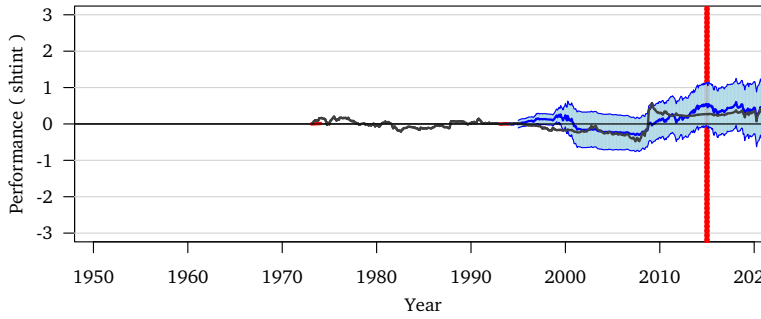

Figure 18
IS and OOS Predictive Performance of RRZ shtint **(monthly)**

performance of shtint was worse than *all-equity-all-the-time* on $1 investment strategies (either equity-tilted or not); and zero or insignificantly better than *all-equity-all-the-time* on Z-scaled investment strategies. **[E]** Figure 18 shows the performance pattern of shtint: it did well from about mid-2008 to mid-2011. It did well again in the Covid bull market.
Evaluation: Although shtint had only marginal IS performance (and showed no advantage in our investment strategies), we suggest actively watching its performance forward.

3.1.19 Y: Yu 2011 Abstract: *[Y] provides evidence that portfolio disagreement measured bottom-up from individual-stock analyst forecast dispersion...is negatively related to ex post expected market return. Contemporaneously, an increase in market disagreement manifests as a drop in discount rate. These findings are consistent with asset pricing theory incorporating belief dispersion.*
Variable: disag is the dispersion of earnings-per-share long-term growth rate forecasts by analysts from the I/B/E/S database, value-weighted across stocks.
Performance: **[A]** Table 2 can confirm the negative and statistically significant IS coefficient of disag in the original sample period (–2005). However, the IS T extended forward to 2021 drops to only –1.14. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of +0.33. The model is unstable. The coefficient turns from negative to positive in our second half (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of disag is negative. **[D]** disag performed worse than *all-equity-all-the-time* on the untilted $1-unscaled investment strategy, but better than *all-equity-all-the-time* on the other three, though never in a statistically significant manner.) **[E]** Figure 19 shows that after good performance from mid-2008 to early 2009 (the early Great Recession), disag has performed poorly.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

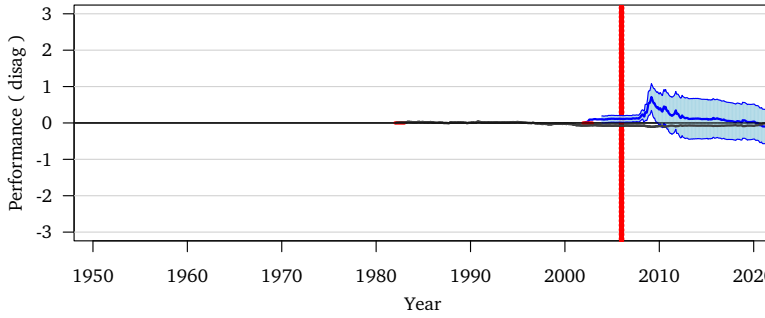

Figure 19
IS and OOS Predictive Performance of Y disag **(monthly)**

Evaluation: We dismiss disag as a useful predictor of equity premiums, based on its poor IS performance (and, further, poor OOSCT performance). Presumably, if the evidence in Yu 2011 was consistent with "a role for market disagreement," the extended evidence should now be viewed as somewhat less consistent

3.2 Quarterly variables

3.2.1 AMP: Atanasov, Møller, and Priestley 2020 Abstract: *[AMP] introduce a novel consumption-based variable, cyclical consumption, and examine its predictive properties for stock returns. Future expected stock returns are high (low) when aggregate consumption falls (rises) relative to its trend and marginal utility from current consumption is high (low). [They] show that the empirical evidence ties consumption decisions of agents to time-variation in returns in a manner consistent with asset pricing models based on external habit formation.*

Variable: The key variable, pce, measures NIPA seasonally adjusted consumption on nondurables and services, provided by the Bureau of Economic Analysis, relative to a trend that is identified by using a filter.

Performance: **[A]** Table 2 can confirm the negative and statistically significant IS coefficient of pce in the original sample period (–2017), with a T of –3.35. With more data to 2021, it strengthens to –3.70. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is still statistically significant. **[C]** Table 3 shows that the OOSCT R^2 of pce is negative.[22] Still, the IS performance was so strong and the OOSCT underperformance so modest that the conjoint p(IS,OOSCT) statistic suggests a solid predictive

[22] In the original paper, the authors began OOS prediction in 1980. This avoided the first 7 years of poor OOS performance in our sample. It was enough to keep pce out of the red zone, though not enough to show meaningfully positive OOS performance (much less with statistical significance). Further unreported investigation shows that our OOS starting forecasting quarter was particularly unfortunate for pce. The OOS performance turns positive with later starting points, though not statistically significantly so.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

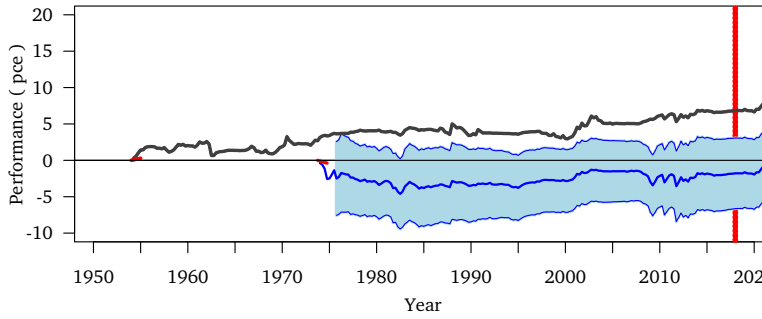

Figure 20
IS and OOS Predictive Performance of AMP pce **(quarterly)**

relationship for pce. **[D]** The investment performance of pce was poor in three of four strategies, and insignificant in the fourth. **[E]** Figure 20 shows that the predictive performance of pce was quite good and consistent in-sample (IS). Since about 1975, the IS performance has still been good but more modest. In contrast, the OOSCT performance was poor for the first ten OOS years from 1973 to 1983. It has improved slowly thereafter. The red line shows that the variable has performed well since publication. It predicted well in the Covid bull market.

Evaluation: Although pce had no useful OOSCT performance and was not of help in our investment strategies, its good IS performance suggests actively watching its performance forward.

3.2.2 BY: Belo and Yu 2013 Abstract: *[BY find that] high rates of government investment in public sector capital forecast high risk premiums.... This result is in sharp contrast with the well-documented negative relationship between the private sector investment rate and risk premiums. To explain the empirical findings, [BY] extend the neoclassical q-theory model of investment and specify public sector capital as an additional input in the firm's technology. [They] show that the model can quantitatively replicate the empirical facts with reasonable parameter values if public sector capital increases the marginal productivity of private inputs.* Naturally, their finding has a strong policy implication, in that it suggests that governments may want to tax and invest more in infrastructure on the margin.

Variable: Their key variable, govik, measures government investment (in contrast to i/k described later, which measures corporate investment). Their original paper's Figure 1 also shows that govik peaked in 1950, then declined until 1982, increased sharply during the Reagan years, then stayed constant, and finally declined again from 2002 to 2010.

Performance: **[A]** Table 2 can confirm the (small) positive and statistically significant IS coefficient of govik in the original sample period (–2010). However, the IS T extended forward to 2021 drops to only 1.46. The model

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

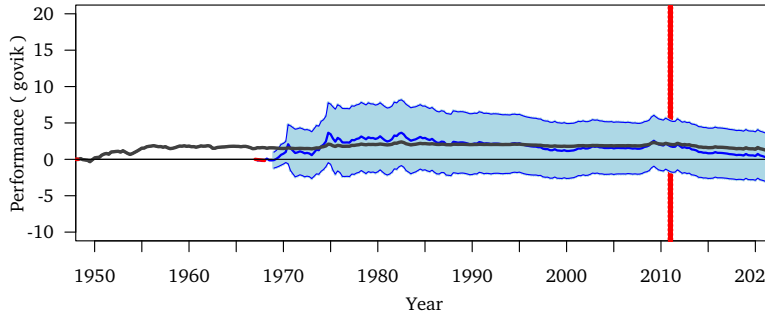

Figure 21
IS and OOS Predictive Performance of BY govik **(quarterly)**

is unstable. The coefficient has turned from positive to negative in our second half (Table A.1). **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is no longer statistically significant. The same coefficient instability afflicts our own homologous regression, too. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of govik is positive, although it is tiny. **[D]** The investment performance of govik was poor. Not only did govik not beat *all-equity-all-the-time*, it even lost money in all but the unscaled equity-tilted strategy.) **[E]** Figure 21 shows that all of the good IS performance was due to early performance. Since about 1960, govik has not predicted well IS. The OOSCT performance had some good predictions, specifically in the early 1970s and again during the oil-crisis from 1973 to 1974, but govik has underperformed ever since.

Evaluation: We dismiss govik as a useful predictor of equity premiums, based on modest IS performance (especially after 1955) and insignificant OOSCT performance. Presumably, if the evidence in Belo and Yu 2013 was consistent with "a role for useful government infrastructure investment," the extended evidence should now be viewed as somewhat less consistent.

3.2.3 CGP: Chava, Gallmeyer, and Park 2015 Abstract: *[CGP analyze] U.S. stock return predictability using a measure of credit standards ('Standards') derived from the Federal Reserve Board's Senior Loan Officer Opinion Survey on Bank Lending Practices. Standards is a strong predictor of stock returns at a business cycle frequency, especially in the post-1990 data period. Empirically, a tightening of Standards predicts lower future stock returns. Standards perform well both in-sample and out-of-sample and is robust to a host of consistency checks. Standards captures stock return predictability at a business cycle frequency and is driven primarily by the ability of Standards to predict cash flow news.*

Variable: crdstd is as obtained from the Fed survey data.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

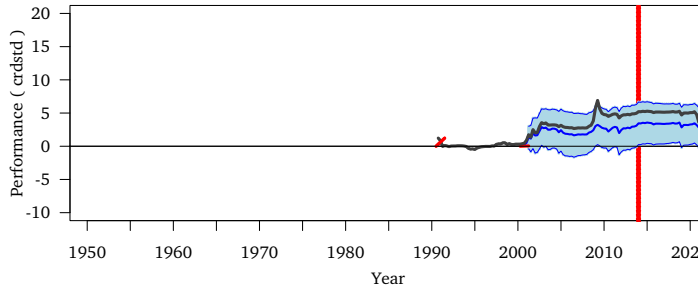

Figure 22
IS and OOS Predictive Performance of CGP crdstd **(quarterly)**

Performance: **[A]** Table 2 can confirm the negative and statistically significant IS coefficient of crdstd in the original sample period (–2013). However, extended forward to 2021, the IS T diminishes to only –1.47. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is marginally statistically significant, with a somewhat better T of –1.67. **[C]** Table 3 shows that the OOSCT R^2 of crdstd is positive, with a strong R^2 of 4.54%. The conjoint statistic strongly suggests a predictive relationship for crdstd. **[D]** crdstd underperformed *all-equity-all-the-time* in the simplest strategy, but outperformed *all-equity-all-the-time* in the other three (though never statistically significantly so). **[E]** Figure 22 shows that crdstd had great performance early on — predicting well from 2000 to mid-2002. Since 2003, crdstd performance has been unremarkable, with a short and mild temporary spike around the time of the Great Recession (predicting Q1-Q2 2009).
Evaluation: Optimistic credit standards predicted poor stock market returns. The variable had good OOSCT performance and was among the best performers on our investment strategies. Furthermore, Table 5 below will show that crdstd also would have helped risk-averse investors achieve higher certainty equivalence. Tempering our enthusiasm, crdstd's in-sample T-statistic as of 2021 is only –1.67. We are further concerned that practically all its good performance originates from its first four years in the sample. However, we consider crdstd worth watching.

3.3 Annual variables

3.3.1 CGMS: Colacito et al. 2016 Abstract: *[CGMS] document that the first and third cross-sectional moments of the distribution of GDP growth rates made by professional forecasters can predict equity excess returns, a finding that is robust to controlling for a large set of well-established predictive factors...time-varying skewness in the distribution of expected growth prospects in an otherwise standard endowment economy can substantially increase the model-implied equity Sharpe ratios, and produce a large amount of fluctuation in equity risk premiums.*

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Variable: CGMS kindly worked with us to isolate the cause for the difference between their data series and our own recalculation. The principal reason is that the data (we repurchased from the vendor) was not the data that was described and used in the original CGMS paper.
Performance: **[A]** We cannot confirm the significant IS coefficient of skew with the vendor data as described in the paper. Our own skew calculation shows no useful predictive ability. (We therefore also include skew only sporadically in later tables.)
Evaluation: We dismiss skew as a useful predictor of equity premiums, because we could not reconstruct the variable as described in the paper.

3.3.2 HHT: Hirshleifer, Hou, and Teoh 2009 Abstract: *[HHT] examine whether the firm-level accrual and cash flow effects extend to the aggregate stock market. In sharp contrast to previous firm-level findings, aggregate accruals is a strong positive time series predictor of aggregate stock returns, and cash flows is a negative predictor...These findings suggest that innovations in accruals and cash flows contain information about changes in discount rates, or that firms manage earnings in response to marketwide undervaluation.*
Variable: HHT introduce two variables: cfacc and accrual. The latter is the difference between earnings and cash flows. HHT use these variables only on annual frequency. For our purposes, it is important to recognize that the two variables are reported by corporations only a few months *after* the closing of their fiscal years. (Our Jan-to-Dec numbers assume no reasonable reporting lag.) Ergo, our focus are on the Jul-to-Jun numbers reported below, which are the only investable ones.

The Accruals Component (accrual)
Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of accrual in the original sample period (–2005), with a T of 2.75. Extended forward, accrual weakens (T of 2.40). **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is still statistically significant. Table 3 shows that this also holds in our extended sample (–2021) and especially in our Jul-Jun mid-years. **[C]** Table 3 shows that the OOSCT R^2 of accrual is positive. The conjoint statistic strongly suggests a predictive relationship for accrual. **[D]** accrual is the first variable that *statistically significantly (at the 9% level)* outperformed *all-equity-all-the-time* in any one of our investment strategies (Jul-Jun, equity-tilted, Z-scaled) at a return of 8%/year, vs. *all-equity-all-the-time* with 3.1%/year. accrual underperformed *all-equity-all-the-time* in $1 investment strategies. **[E]** Figures 23 and 24 show why accrual performed so well. Figure 23 shows that aggregate accruals were consistently flat, with two stark exceptions: 1973–1974 (conservative) and 1999–2001 (aggressive). The former occurred before our OOS analysis begins. Figure 24 shows that the latter was a great call.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

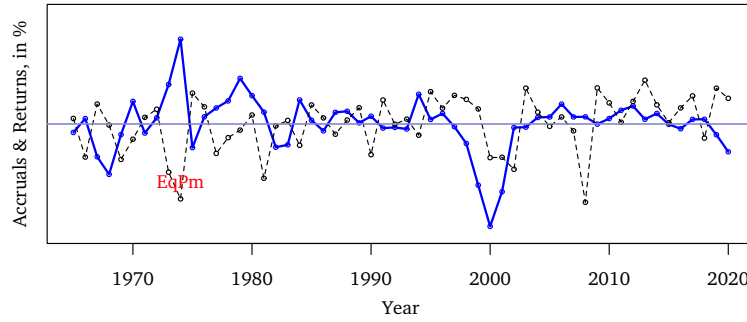

Figure 23
Time-Series of Accruals (accrul) **and Equity Premia**

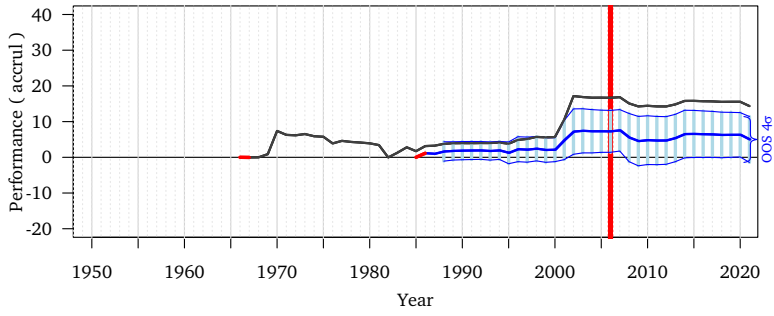

Figure 24
IS and OOS Predictive Performance of HHT accrul **(annual/jun)**

The market declined greatly in 2000–2002, following the dot-com years. In "ordinary years," aggregate accruals were unremarkable. They barely budged. Evaluation: accrul is a difficult variable to assess primarily due to its episodic performance.[23]

One can hold the view that managers' over-optimism or over-pessimism anticipated the (opposite) reversal of investors' sentiment in a particular kind of market exuberance that was followed by its predictable collapse. (Of course, corporate managers aggressively increasing their accruals were not sufficiently aware that they could have done even better opening hedge funds.)

Or one can hold the view that the 1999–2001 event was too singular a period to make it likely that accrul will help again predict equity premia in the future. (Table 5 will also show that a risk-averse investor would not want to use accrul for timing.)

[23] Non-behavioral researchers would secondarily wonder about the dynamics of the widely but not universally acknowledged irrational exuberance and the accounting conventions of managers of the era.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Realistically, researchers have to rely on their perspectives, at least until there will be a few more episodes in which accrul increases or decreases dramatically.

The Cash Flow Component (cfacc)
Performance: **[A]** Table 2 can confirm the negative and statistically significant IS coefficient of cfacc in the original sample period (–2005). However, one problem is that cfacc performs well only if there is no reporting lag (Jan-Dec but not Jul-Jun). With a reporting lag, the IS T falls from –3.14 (Dec) to –1.47 (Jun) in Table 2. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is still statistically significant. Again, with a reporting lag, the IS T falls from –3.09 (Dec) to –1.44 (Jun). **[C]** The OOSCT R^2 of cfacc are strongly positive. The conjoint statistic strongly suggests a predictive relationship for cfacc. **[D]** cfacc outperformed *all-equity-all-the-time* except in the simplest untilted $1 strategy. **[E]** We skip this figure (and further evaluation) because accrul is the better variable.

3.3.3 MR: Møller and Rangvid 2015 Abstract: *[MR] show that macroeconomic growth at the end of the year (fourth quarter or December) strongly influences expected returns on risky financial assets, whereas economic growth during the rest of the year does not. We find this pattern for many different asset classes, across different time periods, and for US and international data.*

It is worth noting that the paper's perspective that the fourth quarter data is special was motivated by Jagannathan and Wang 2007.
Variable: MR introduce two variables: gpce and gip, both based on OECD data. The former is the growth rate in personal consumption expenditures, the latter is the growth rate in industrial production. The variables are available on a quarterly basis, but Møller and Rangvid 2015 use them only on an annual basis, presumably due to the special fourth-quarter perspective. Therefore, we kept their frequency choice.

We are also not certain about the release timing of the MR variables and whether it would be the (more easily investible) early information release or the actual economic performance (possibly known by investors but not researchers) that should predict stock returns. For example, an OECD revision spreadsheet suggests that for Q4 2000 (Oct, Nov, Dec), their monthly IIP series (itself having a monthly time-series variance of about 1%) was revised as follows:

First Release	+2mos	+1mo	+9mo	+21mo	Latest
1.40	1.73	1.40	1.55	1.78	2.79

The Growth Rate in Personal Consumption Expenditures (gpce)
Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of gpce in the original sample period (–2009), with a

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

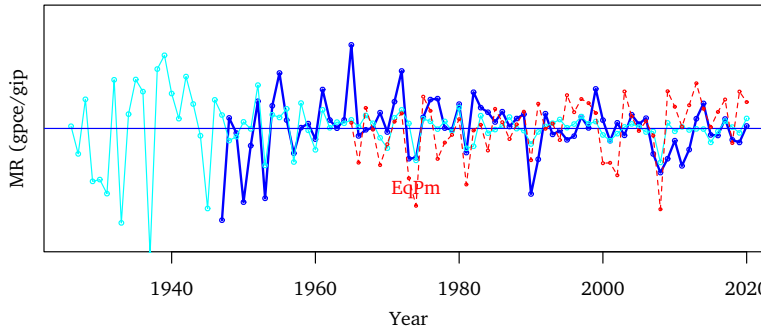

Figure 25
Time-Series of Personal Expenditures Growth (gpce) and Equity Premia

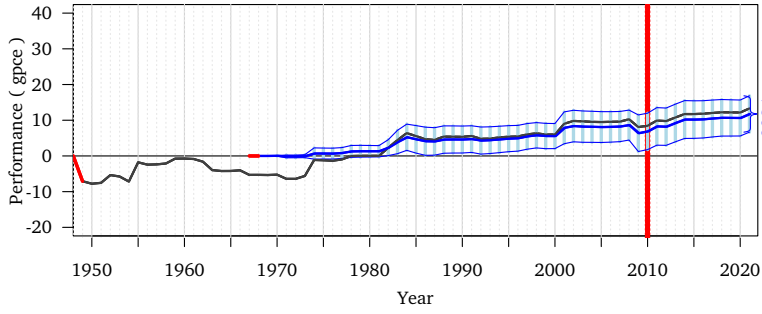

Figure 26
IS and OOS Predictive Performance of MR gpce (annual/jun)

T of –4.54(!). In our extended sample (–2021), the IS coefficient even improves its T to –5.13(!). **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is still statistically significant, though with a more modest T statistic of –3.61. With a 6-month reporting delay, it falls to –1.94. **[C]** Table 3 shows that the OOSCT R^2 of gpce is positive. The conjoint p(IS,OOSCT) statistic strongly suggests a predictive relationship for gpce. **[D]** With a reporting delay, gpce outperforms *all-equity-all-the-time* in all cases. In the tilted and Z-scaled version, it does so in a statistically significant manner on a one-sided test at the 10% level with a T-statistic of 1.51. **[E]** Figure 25 plots the time-series of gpce. There are no obvious patterns. In recessions, gpce declines. Figure 26 shows that gpce had good performance beginning with its 1974 prediction for 1975, with the only misprediction being its wrong 2007 call about the 2008 Great Recession bear market.

Evaluation: The gpce (fourth-quarter) variable was a very good equity-premium predictor in our sample: Years in which consumers spent more are followed by bear stock markets. gpce also performed well after the original authors' sample period had ended. (One caveat is that the reader must assess

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

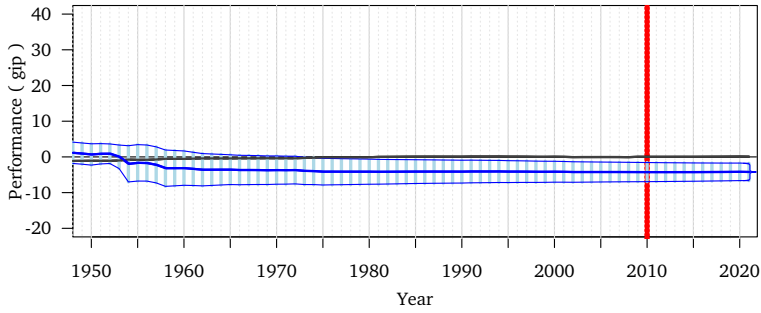

Figure 27
IS and OOS Predictive Performance of MR gip (annual/dec)

whether the ex-post choice of the fourth quarter raises data-snooping as a concern — though Jagannathan and Wang 2007 suggest some good reasons why this should be so.)

The Growth Rate in Industrial Productions (gip)

Performance: **[A]** Table 2 can confirm the strong negative and statistically significant IS coefficient of gip in the original sample period (–2009). However, unlike gpce, which has only been available since 1947, gip has been available since 1926. Once gip's sample is extended backwards to 1926, gip no longer predicts well even IS. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with IS Ts of –0.09 and –0.17 in calendar year and delayed samples. This is due to backward extension. Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of gip is negative. **[D]** The investment performance of gip was poor.)

Evaluation: We dismiss gip as a useful predictor of equity premiums, based on poor IS and OOSCT performance in the full sample extending backward. However, again, the poor performance of gip is less driven by the fact that our sample extends forward to 2021 and more by the fact that gip (*unlike* gpce) has been available since 1926. Remaining consistent with our treatment of other variables, our paper uses the entire data series. The authors focus on the shared sample beginning in 1947. After 1947, gip had better performance and behaved generally very much like gpce. **[E]** Because gpce and gip are so similar, we did not graph gip.

3.3.4 PST: Piazzesi, Schneider, and Tuzel 2007 Abstract: *[PST] consider a consumption-based asset pricing model where housing is explicitly modeled both as an asset and as a consumption good...the model predicts that the housing share can be used to forecast excess returns on stocks. [They] document that this indeed true in the data. The presence of composition risk also implies that the riskless rate is low which further helps the model improve on the standard CCAPM.*

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

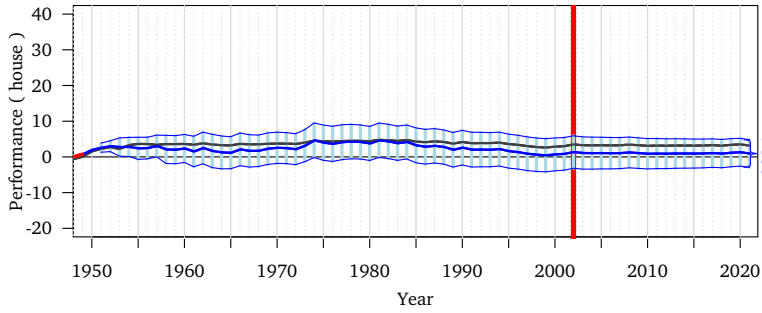

Figure 28
IS and OOS Predictive Performance of PST house **(annual/dec)**

Variable: house is a measure of the dollar amount spent on rent or estimates of how much owners would rent their houses for. (The paper shows that effective rents as a fraction of income declined from 1930s to the 1980s and then stabilized.) The regressions work with overlapping 5-year periods.
Performance: **[A]** Table 2 can confirm the strong positive and statistically significant IS coefficient of house in the original sample period (–2001), with a T of 2.64. Extended to 2021, the T statistic increases to 2.97. However, the model is unstable. Table A.1 shows that the T was 3.65 in the first half of the authors' sample and –2.65 in the second half. Extending the sample backwards by one more five-year period (1926 instead of 1931) also eliminates the positive significance of house. **[B]** In our homologous specification (–2021) in Table 3, the IS coefficient is not statistically significant, with an IS T of 0.96 without and 0.65 with reporting delay. The model is unstable (in both calendar-year and delayed regressions), with positive coefficients in the first halves and a negative coefficient in the second halves (Table A.1). Thus, with poor IS performance, further OOS investigation seems largely unwarranted. (**[C]** Table 3 shows that the OOSCT R^2 of house is positive. **[D]** The investment performance of house was poor. The non-delayed non-equity-tilted strategies not only did not beat *all-equity-all-the-time*, they lost money in absolute terms. The delayed strategies merely underperformed *all-equity-all-the-time*.) **[E]** Figure 28 shows that the performance of house was largely non-descript. It may have performed better before 1980 and worse thereafter.
Evaluation: We dismiss house as a useful predictor of equity premiums, based primarily on its poor IS performance. Presumably, if the evidence in Piazzesi, Schneider, and Tuzel 2007 was consistent with "a CCAPM with housing," the extended evidence should now be viewed as somewhat less consistent

3.4 Revisiting Variables in Goyal and Welch 2008

Since Goyal and Welch 2008 was published, 15 years have passed. Our paper is a good opportunity to revisit variables already examined therein, too.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

[A] We do not look at IS performance in original author sample periods. **[B]** Most variables in Goyal and Welch 2008 were of monthly frequency. Remarkably, the dividend ratios (which were the variables that spawned the equity prediction literature) have continued to perform poorly. (Goyal and Welch 2003 and Van Binsbergen and Koijen 2010 explored this further.) Only one monthly variable (the Treasury-bill yield, tby) shows good IS statistical significance, with a T of –1.82. The long-term Treasury yield (lty) is close with a T of –1.57. **[C]** With good OOSCT performance, both tby and lty exceed the conjoint p(IS,OOSCT) statistic for performance expected under the null hypothesis. **[D]** Nevertheless, in our four investment strategies, neither earned higher rates of return than the *all-equity-all-the-time* alternative. Surprisingly, both variables' investment strategies even *lost* money in absolute terms! The yield is of course a negative component of the variable being predicted, too. **[E]** Figures 29 and 30 show that most of the good performance of tby and lty dates back to the 1974–1975 oil-shock episode. From 1984 to 2008, their predictions deteriorated but then recovered. Both had good COVID performance. Evaluation: Neither of the two monthly variables, both

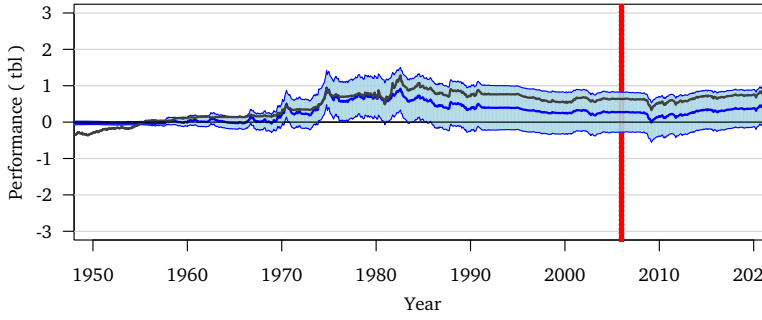

Figure 29
IS and OOS Predictive Performance of CA tby **(monthly)**

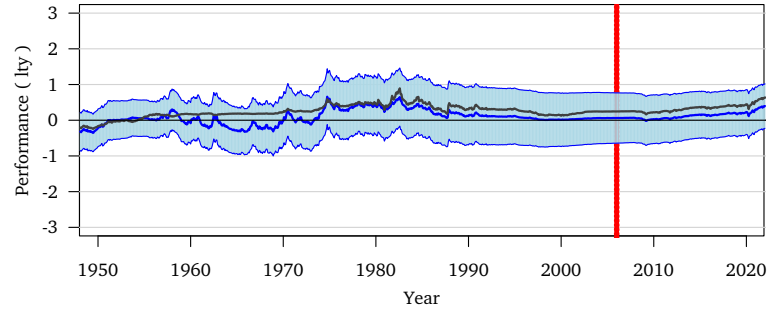

Figure 30
IS and OOS Predictive Performance of FF lty **(monthly)**

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

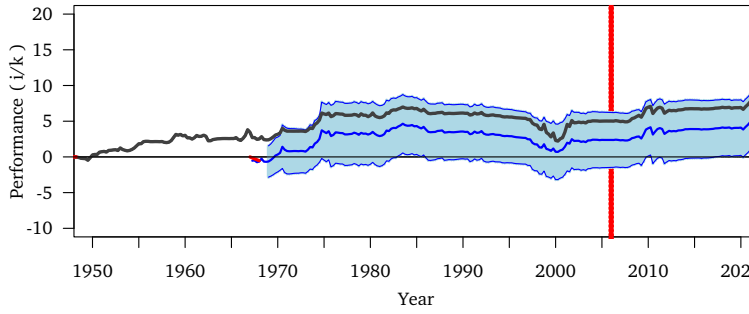

Figure 31
IS and OOS Predictive Performance of Co i/k (quarterly)

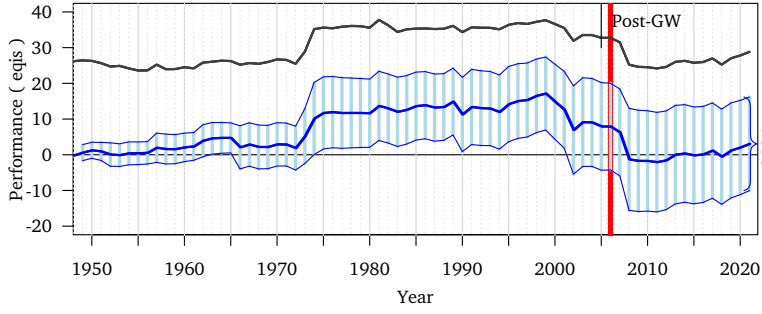

Figure 32
IS and OOS Predictive Performance of BW eqis (annual/dec)

fixed-income yield based, showed great predictive power. Both seem worth watching, even though they delivered very poor investment results.

Goyal and Welch 2008 examined only two quarterly variables, i/k and cay. cay performs poorly both IS and OOS, as well as for investment purposes. i/k however has very good IS and OOSCT performance. Figure 31 shows that i/k's predictions have steadily improved since Goyal and Welch 2008. (The low point occurred in 2000.) We would recommend considering i/k as a predictor, even though i/k did not help an investor outperform *all-equity-all-the-time* on any of our four investment strategies.

Goyal and Welch 2008 examined only one annual variable, eqis. It still performs well IS and OOS. This is even though Figure 32 suggests nondescript OOSCT performance.[24] The performance of eqis was strongest during the 1970s oil crisis, and weakest from 1998 to 2009. Since 2009, it has been largely nondescript.

[24] Caveat: The figure is based on residuals and thus its OOSCT line corresponds well to R^2, while the statistical significance is based on the MSE t-statistic. Usually, the two offer similar inference, except for eqis.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Not shown in the tables, since Goyal and Welch 2008, the OOSCT performances of 13 of the 17 variables have further deteriorated, the performances of three variables (ltr, lty, and tby) have improved slightly, and only i/k has improved visibly (Figure 31). More data did not help the original GW set of variables. Instead, the overall tally has become worse than one would have expected by chance.

4. Risk-Averse Investors' Certainty Equivalence

Campbell and Thompson 2008 also suggested considering the predictors from the perspective of a risk-averse investor. We now closely follow their guidance.

First, we determine the recommended equity investment for a quadratic risk-averse investor with $\gamma = 5$, who believes the *annualized* prevailing stock return variance ($\widehat{V_t(R)}$) will continue, and who further believes that the *annualized* expected rate of return in the stock market is given by her modeled beliefs as $\widehat{E_t(R)}$. Her optimal stock market tilt is then

$$w_{\mathrm{EQ},t} = \frac{\widehat{E_t(R)}}{\gamma \cdot \widehat{V_t(R)}},$$

further bounded to be positive and not to exceed 150%. This is again as suggested by Campbell and Thompson 2008 and enhanced by Löffler 2022.

As alternatives, we consider three different variable-unconditional strategies: First, one in which a (similarly risk-averse) investor believes that the expected rate of return on the stock market is the historical average rate of return in time. Second, one in which the investor holds about 50% in equities. This is, presumably, a risk-averse investor would not hold 100% in equities. With a gamma of 5, this could be consistent with expecting an annual rate of return similar to the annual standard deviation. Third, one in which the investor is more aggressive and holds *all-equity-all-the-time*.

Each investment strategy would then have offered a time-series of return realizations. To compare them from the perspective of this risk-averse investor, the principal metric is

$$\Delta \mathrm{UTIL} = \left(\hat{\mu}_{\mathrm{Mdl}} - \frac{\gamma}{2}\hat{\sigma}^2_{\mathrm{Mdl}}\right) - \left(\hat{\mu}_{\mathrm{Unc}} - \frac{\gamma}{2}\hat{\sigma}^2_{\mathrm{Unc}}\right),$$

where $\hat{\mu}_{\mathrm{Mdl}}$ and $\hat{\sigma}^2_{\mathrm{Mdl}}$ are the realized mean and variance of the rate of return on the portfolio of the conditional strategy; and $\hat{\mu}_{\mathrm{Unc}}$ and $\hat{\sigma}^2_{\mathrm{Unc}}$ are the corresponding statistics for the unconditional non-variable-based strategy. We calculate the statistical significance of the utility difference following the procedure in DeMiguel, Garlappi, and Uppal 2009, Footnote 16.

We focus only on the most promising variables from the previous tables.

Table 5 shows that the standout performers for an investor relative to the historical average equity premium would have been the quarterly crdstd, (2.52% to 5.82% CEV per annum), and the monthly shtint (1.61% to 3.92%).

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 5
Risk-averse investor certainty equivalence value performance

		Δu(Mdl−Unc)		Δu(Mdl −50% Eq)		Δu(Mdl −100% Eq)	
Freq	Vrbl	CEV	T-stat	CEV	T-stat	CEV	T-stat
M	wtexas	0.44	(0.64)	–0.33	(–0.30)✗	–0.46	(–0.52)✗
M	tchi	1.56	(1.69)	0.40	(0.37)	0.92	(0.66)
M	shtint	3.36	(2.31)✔	2.25	(1.22)	1.61	(1.12)
M	tby	1.24	(1.53)	0.47	(0.50)	0.34	(0.33)
M	lty	0.99	(1.05)	0.22	(0.24)	0.10	(0.09)
M	infl	0.82	(1.94)	0.06	(0.06)	–0.07	(–0.09)✗
Q	crdstd	5.82	(2.71)✔	1.74	(0.79)	3.30	(1.20)
Q	avgcor	1.19	(2.00)✔	–0.45	(–0.35)✗	–0.07	(–0.10)✗
Q	i/k	2.16	(1.40)	–0.33	(–0.25)✗	0.96	(0.64)
D	accrul	0.81	(0.76)	–3.10	(–1.47)	–3.18	(–1.42)✗
D	gpce	0.62	(0.46)	–2.92	(–1.29)✗	–1.63	(–0.98)✗
D	gip	0.62	(1.62)	–2.97	(–1.68)✗	–2.05	(–1.88)✗
D	eqis	1.26	(1.25)	–2.06	(–1.02)✗	–1.31	(–0.95)✗
J	accrul	0.20	(0.20)	–1.62	(–0.86)✗	–2.13	(–1.04)✗
J	gpce	1.44	(2.22)	–2.13	(–1.14)✗	–0.48	(–0.42)✗

Explanations: Papers, variables, and sample periods are defined in Table 1. The table describes the utility improvement of selected variables for a quadratic risk-averse investor with risk-aversion parameter 5 and investment limits of 0 and 1.5 times equity, just as in Campbell and Thompson 2008. The "Δu(Mdl−Unc)" column computes a prevailing mean return ($\widehat{E(R)}$) for both the conditional model (using a variable) and the prevailing-unconditional-mean model. The variance ($\widehat{V(R)}$) is the rolling market variance, calculated using monthly data over the last 5 years. It then calculates both investment strategies' weights as $w_{\text{Equity}} = \widehat{E(R)}/[\gamma \cdot \widehat{V(R)}]$, financed by the short rate. The difference in realized returns between the two investment strategy yields a time-series of realized portfolio returns (R_{pt}). The reported utility difference is then $E(R_{pt}) - \gamma/2 \cdot V(R_{pt})$. The two other comparisons are relative to a strategy always investing 50% and 100% in equity, independent of the prevailing historical prevailing-mean model. All certainty equivalence values (CEVs) are annualized.

Interpretation: Some variables improved this risk-averse utility relative to an prevailing-mean model, but none of the variables improved it relative to a simple 50%-equity or to a 100% buy-and-hold strategy in a statistically significant manner. However, shtint came close. Quarterly and annual variables — generally better predictors for a risk-neutral investor than monthly variables — would not have offered similar attractiveness for a risk-averse investor. However, crdstd performs very well.

Because both variables were not particularly good predictors of expected rates of return in Table 3, a large part of their performance is likely due to their intelligence about prevailing risk. Other variables, like i/k, that predicted expected return better would still have allowed an investor to improve utility, but would not have done so in a statistically significant manner. Remarkably, despite their good performances for a risk-neutral investor, some other variables, such as accrul and gpce would not have helped a risk-averse investor relative to the simple *all-equity-all-the-time*. The additional realized variance (because of aggressive positions in high-volatility markets) would have been bad enough to negate the advantage of the improved mean prediction.

5. Assessment of Tally of Performance

As noted earlier, our visual summary in **Table 6** sacrifices detail and nuance but makes it easy to quickly assess the relative performance of different variables on different criteria.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 6
Overview of a reasonable qualitative assessment of the findings

Panel A: Monthly-frequency variables

		IS performance					Other performance		
Table 1		Table 2			Table 3	Tbl A.1	Tbl 3		Tbl 4
Ppr	Var	Same	Forw	F/B	F/B	Halves	OOSCT	IS,OOSCT	InvZLE
BH	vp	✔	✔	✔	✗	..	✗	✗	☒✗✗✗
BPS	impvar	✔	✗	✗	✗	..	✗	✗	☒☒✗✗
BTZ	vrp	✔	✗	✗	✗	·✗	·	✗	✗✗✗✗
CEP	lzrt	✔	✗	✗	✗	·✗	·	·	☒☒✗✗
CP	ogap	✔	✔	†✗	†✗	..	†✗	†✗	✗✗✗✗
DJM	wtexas	✔	✗	✗	✗	..	✔	·	✗✗✗✗
HJTZ	sntm	✔	✔	✔	✗	..	✗	·	✗✗✗✗
JT	ndrbl	✔	✔	✔	✗	..	✗	✗	✗✗☒☒
JZZ	skvw	✔	✗	✗	✗	·✗	✗	✗	☒☒✗✗
KJ	tail	✔	✔	†✗	†✗	·✗	†✗	†✗	✗✗✗✗
KP	fbm	✔	✔	✔	✗	..	✗	✗	✗✗✗✗
LY	dtoy	✔	✗	✗	✗	✗✗	✗	✗	☒✗☒☒
LY	dtoat	✔	✔	✔	✗	..	✗	✗	☒☒☒✗
$Maio_{(13)}$	ygap	✔	✗	✗	✗	..	✗	✗	☒☒✗✗
$Maio_{(16)}$	rdsp	✔	✗	✗	✗	·✗	✗	✗	☒☒☒✗
Mrtn	rsvix	✔	✔	✔	✗	·✗	✗	✗	✗✗✗✗
NRTZ	tchi	✔	✔	✔	✗	..	✔	✔	✗☒··
PW	avgcor	✔	✔	†✗	†✗	..	†✗	†✗	✗✗✗✗
RRZ	shtint	✔	✔	✔	✗	..	✔	✔	✗·✗·
Y (pre-2008)	disag	✔	✗	✗	✗	·✗	✗	✗	✗···
BMRR	ntis	n/a	n/a	n/a	✗	n/a·	✗	✗	✗✗☒☒
Cmpl	tby	n/a	n/a	n/a	✔	n/a·	✔	✔	☒☒☒✗
CSa	d/p	n/a	n/a	n/a	✗	n/a·	✗	✗	☒☒☒✗
CSb	d/y	n/a	n/a	n/a	✗	n/a·	✗	✗	☒☒☒✗
CSc	e/p	n/a	n/a	n/a	✗	n/a·	✗	✗	☒✗✗✗
CSd	d/e	n/a	n/a	n/a	✗	n/a✗	✗	✗	☒☒☒☒
CSe	svar	n/a	n/a	n/a	✗	n/a✗	✗	✗	☒✗✗✗
FFa	lty	n/a	n/a	n/a	✗	n/a·	✔	✔	☒✗☒☒
FFb	ltr	n/a	n/a	n/a	✗	n/a·	✗	✗	✗·✗·
FFc	tms	n/a	n/a	n/a	✗	n/a·	·	✗	✗✗✗·
FFd	dfy	n/a	n/a	n/a	✗	n/a·	✗	✗	☒✗✗✗
FFe	dfr	n/a	n/a	n/a	✗	n/a·	✗	✗	✗✗✗✗
FS	infl	n/a	n/a	n/a	✗	n/a·	✔	·	✗·☒✗
KS	b/m	n/a	n/a	n/a	✗	n/a✗	✗	✗	☒☒☒✗

Panel B: Quarterly-frequency variables

		IS performance					Other performance		
Table 1		Table 2			Table 3	Tbl A.1	Tbl 3		Tbl 4
Ppr	Var	Same	Forw	F/B	F/B	Halves	OOSCT	IS,OOSCT	InvZLE
AMP	pce	✔	✔	✔	✔	..	✗	✔	✗✗✗✗
BY	govik	✔	·	✗	✗	✗✗	·	·	☒☒☒☒
CGP	crdstd	✔	·	✗	✗	..	✔	✔	✗·✗·
Crn	i/k	n/a	n/a	n/a	✔	n/a·	✔	✔	✗✗☒☒
LL	cay	n/a	n/a	n/a	✗	n/a✗	✗	✗	✗✗☒✗

In general, lower-frequency variables tended to predict the log-equity premium better than higher-frequency variables. Disappointing performance was more common in monthly variables. Quarterly and annual variables did much better.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table 6
Continued

Panel C: Annual-frequency variables, January to December

		IS performance					Other performance		
Table 1		Table 2			Table 3	Tbl A.1	Tbl 3		Tbl 4
Ppr	Var	Same	Forw	F/B	F/B	Halves	OOSCT	IS,OOSCT	InvZLE
CGMS	skew	✗	✗	✗	✗	✗✗	✗	✗	✗✗✗✗
HHT	accrul	✔	✔	✔	✔	··	✔	✔	✗·✗·
HHT	cfacc	✔	✔	✔	✔	··	✔	✔	✗···
MR	gpce	✔	✔	✔	✔	··	✔	✔	✗·☒·
MR	gip	✔	✔	✗	✗	·✗	✗	✗	✗✗☒✔
PST	house	✔	✔	✗	✗	✗✗	✔	·	☒☒✗☒
BW	eqis	n/a	n/a	n/a	✔	n/a·	✔	✔	✗✗··

Panel D: Annual-frequency variables, July to June

		IS performance					Other performance		
Table 1		Table 2			Table 3	Tbl A.1	Tbl 3		Tbl 4
Ppr	Var	Same	Forw	F/B	F/B	Halves	OOSCT	IS,OOSCT	InvZLE
CGMS	skew	✗	✗	✗	✗	✗·	✗	✗	·✗·✗
HHT	accrul	✔	✔	✔	✔	··	✔	✔	✗·✗✔
HHT	cfacc	✗	✗	✗	✗	··	✔	✔	✗✗·✗
MR	gpce	✔	✔	✔	✔	··	✔	✔	···✔
MR	gip	✗	✔	✔	✗	··	✗	✗	✗✗✗✗
PST	house	✔	✗	✗	✗	✗✗	✗	✗	✗✗✗✗
BW	eqis	n/a	n/a	n/a	✗	n/a✗	✗	✔	✗✗✗✗

Explanations: This table is a visual summary of the qualitative results from other tables. Other interpretations are possible, just like researchers can choose to mark different statistical significance levels. For details and caveats, please refer to the earlier tables and their in-text discussions themselves.

Data in the first column characterize our results in an "author-similar" IS regression. The second column describes the same results extending the sample forward to end in 2021. The third column extends the sample forward and backward to the longest available data series. Both columns are ✔'ed only if the (IS) variable remains statistically significant, and ✗'ed out if they lose their in-sample significance. The fourth column uses our specifications (log equity premium, etc.) and extends the sample forward and backward to the longest available data series. The fifth column ✗'s variables that change IS sign in the first and second halves of a sample, in the original authors' sample and in our extended sample. (Variables that are lower or higher significantly in one half but do not change sign are left blank.) The sixth column shows the Campbell-Thompson improved OOSCT R^2. Negative values are ✗'ed out, insignificant positive values are left blank, and significant positive values are ✔'ed. The seventh column shows the joint p(IS,OOSCT) performance — ✔ if significant at the 5% level, ✗ if not even significant at the 10% level. The eighth column shows the in-time investment performance. Variables that underperformed an equivalent buy-and-hold strategy are ✗'ed out (with an extra box if they lost money in absolute terms), variables that outperformed but not statistically significantly so are left blank. Three monthly variables (ogap,tail,avgcor) performed much better when not extended backwards, too, which is marked by a preceding †.

Interpretation: The variable that appeared most appealing on many criteria was accrul. In total, 13 of 46 variables held in the authors' sample forward and did not have negative OOSCT R^2. Monthly variables include: tchi and shtint. The Treasury-yield variables tby, and possibly lty, and tms also did well. Quarterly variables include: pce, crdstd, and i/k. Annual variables include accrul, cfacc, gpce, and eqis. Of all 46 variables with 184 considered investment strategies, only three (accrul, gpce, and gip) outperformed *all-equity-all-the-time* in a statistically significant manner and each in only one of four investment strategies.

Monthly: Of our 20 new monthly-frequency variables (post-GW), only 4 improved on their t-statistic as new data was added. Sixteen variables deteriorated, 9 variables badly enough that they have already lost their IS significance since publication. Three more variables lose their IS significance

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

if we also extend the sample backward. Ten new variables do retain their IS significance. Internet Appendix IA.I describe simulations that suggest that this poor a performance is more consistent with spurious associations to begin with than with any association.

In the homologous IS specification, in which variables have to predict nonoverlapping log-equity premiums at native frequencies, all 20 new monthly variables lose their statistical significance. Nevertheless, 2 variables retain sufficient OOSCT performance to render conjoint IS-OOSCT significance: tchi and avgcor (though only if the inference does not start in 1926, when data becomes available). Moreover, 3 monthly variables do not lose their IS and OOSCT performances if their data is not extended backward beyond World War 2: ogap, tail, and avgcor.

From the pre-Goyal and Welch 2008 monthly variables, two perform well. The Campbell 1987 Treasury-bill yield (tby) and the Fama and French 1989 long-term yield (lty) show good IS statistical significance. Higher fixed-income yields predict lower equity premiums. The two yields also show good OOSCT performance. Thus, they are our best monthly predictors. Unfortunately, even these two variables did not allow any of our four simple investment strategies to outperform *all-equity-all-the-time*. The tby even lost money in absolute terms in all four strategies; lty in three or four strategies.

Quarterly: On a quarterly basis, 2 of 5 new variables showed good IS performance: pce (aggregate consumption to trend, Atanasov, Møller, and Priestley 2020) and i/k (the investment-capital ratio, Cochrane 1991). However, one-third, crdstd (credit standards, Chava, Gallmeyer, and Park 2015) just barely missed good IS performance in our own (rather than the authors') specification. Both i/k and crdstd also showed good OOSCT performance. Thus both had conjoint p(IS,OOSCT) significance levels below 1%.

High investment this quarter predicted poor stock-market returns the following quarter. Interestingly, like gpce, i/k associates more outlays today with lower market performance in the future—almost as if the alternative had been holding back and stockpiling funds today perhaps to allow for more market investment later. For the 13 years from 1975 to 1998, i/k was a poor predictor. In the 23 years since then, i/k has consistently performed well. Thus, it performs better today than it did in Goyal and Welch 2008. Tempering our enthusiasm, Table A.1 shows that its estimated IS coefficient in our sample has declined from –3.74 in our first half to –1.66 in the second half; and i/k could not outperform *all-equity-all-the-time* in any of our four timing strategies. In fact, two of four i/k-based investment strategies lost money even in absolute terms.

In contrast, crdstd did perform well in Z-scaled strategies, yielding an economically, but not statistically, significant return as much as 5% higher than that of *all-equity-all-the-time*. crdstd also would have helped a risk-averse investor.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Annual: On an annual basis, 4 of 7 variables showed both good IS and OOSCT performances: accrual and cfacc (accruals and cash-flow accruals, Hirshleifer, Hou, and Teoh 2009), gpce (year-end economic growth, Møller and Rangvid 2015), and eqis (equity-issuing, Baker and Wurgler 2000). In the 16 simple investment strategies that we considered for these 4 successful variables, the equity-tilted Z-scaled strategies for accrual and gpce beat *all-equity-all-the-time* with *t*-statistics of about 1.3 and 1.5 and Sharpe-ratios of about 0.2. Some other variables (eqis) performed about the same as *all-equity-all-the-time*. The other variables and scenarios still underperformed *all-equity-all-the-time*.

Aggressive corporate accruals reliably predicted low future stock returns. Nevertheless, accrual was a "one-trick pony"—but with a superb trick: It predicted the dot-com aftermath bear market without losing out on the preceding and succeeding bull markets.[25] This is its only strong—and singularly stellar—prediction. Since 2002, accrual has not moved much. Thus, its single outlier performance was enough to at first obtain and subsequently avoid losing its performance in our extended sample. (Incidentally, a risk-averse investor would not have been better off using accrual.)

gpce was a more consistent performer. Since the 1970s, gpce has only made one modest misstep in its predictive ability (which was missing the Great Recession bear market). Otherwise, gpce has been a steady performer. (Incidentally, a risk-averse investor, as defined by Campbell and Thompson 2008, would *not* have been better off using gpce.)

Finally, equity-issuing activity, eqis, a pre-GW variable, also had good IS and OOSCT performance, though not particularly good investment performance.

In light of our data to 2021, we would not dismiss 11 of our 46 variables out of hand: tchi, shtint, tby, lty, tms; pce, crdstd; and i/k; accrual, gpce, and eqis. The choice among them may depend on other criteria, such as how to read the singular prediction of accrual. The strongest variable *in the post-WWII era*, generally, seems to have been gpce. The inference about the other 35 of our 46 variables seems underwhelming.

Annual or quarterly from monthly variables: Some authors also described longer-horizon performance for monthly variables. Moreover, the quarterly, semiannual, and annual variables seemed to have better predictive performance in general. This could have been due to the specific predictor variables or because the predicted variables were longer-horizon (and thus perhaps less noisy) equity premiums. The Internet Appendix IA.VI shows that among the monthly variables, two variables (JT ndrbl and CEP lzrt)

[25] Missing our OOS prediction incept sample need, accrual's prediction in the 1974 bear market would have been its second trick.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

have significantly improved performance on longer-horizon equity predictions. That is, they acquire good homologous-regression IS performance and, in some specifications, also good OOS performance. The other 19 monthly variables remain insignificant. (Among the original GW variables, ltr and eqis are also improving.)

6. A Weighted Prevailing Best Estimator?

Our paper is not focused on finding a better predictor of equity premiums. As noted, our paper's purpose is instead to offer diagnostics for previously described variables.

It would have been difficult to know ex ante what variables to trust and what not to trust. One possible approach would have been to look at statistics of many predictors. Table 7 shows that the "single-best prevailing variable" never offered a positive OOSCT R^2. An average "consensus" combinations—in which variables were weighted according to their prevailing t-statistics and in which there was no assumed information release delay—did perform well with an OOSCT R^2 of 3.68%.

However, Figure 33 shows that this consensus performed well only up to the mid 1970s. Recall that these are sum-squared errors, not R^2, so a horizontal line is not indicative of continued good performance. Instead, a horizontal line is indicative of zero predictive ability. The consensus had three good episodes: 1950-1955 (+1%); 2002-2003 (+0.5%), and 2012 (+0.3%). It generally predicted poorly from 2012-2021 (–0.8%).

We also tried to construct investment strategies analogous to those in Table 4. For example, we took the t-statistic based prevailing linear combination of equity premiums prediction to check whether the current prediction was above or below average. The strategies always performed worse than *all-equity-all-the-time*. (In an ex ante equity-tilted strategy, the combination consensus would always have been long, similar to, for example, Y's disag, which was long in only one month.)

Table 7
Combination predictors OOSCT R^2 (with CT prediction truncation)

	Monthly	+ Quarterly	+ Annual	+ w/delay
Best prevailing predictor	−3.26	−9.21	−17.42	−12.64
Average predictor	−0.05	0.04	0.63*	−2.74
T-weighted predictor	0.10*	0.53**	3.68**	−0.73

Explanations: In each period, we take the prevailing estimators to create one single forecast. We truncate the final prediction. Note that the lower-frequency predictions here also include higher-frequency variables. Thus, the annual predictions also include monthly variables, too.

Interpretation: The T-weighted combination calendar-year forecast (without delay) provided good predictions. (Figure 33 shows that this performance was entirely due to the performance in the first half of our sample.)

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

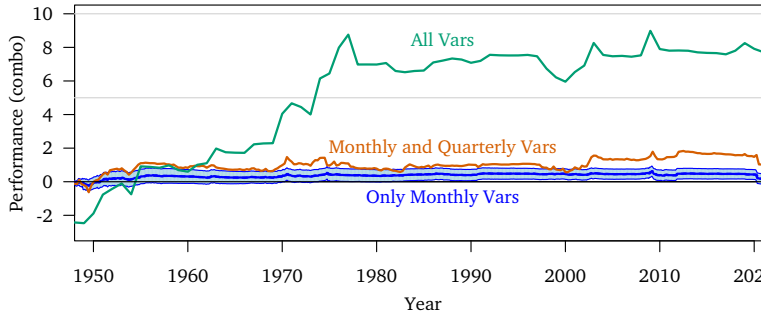

Figure 33
OOSCT predictive performance of the T-weighted combination estimator
Explanations: The blue line represents the performance of the T-weighted predictions using only monthly variables. The red line adds quarterly variables, and the green line adds annual variables. Recall also that these are not R^2 but sum-squared cumulative errors. A horizontal line represents zero predictive ability.
Interpretation: The performance using all variables and predicting on annual bases was excellent up to the mid-1970s. It was nondescript thereafter; that is, the consensus has neither outperformed nor underperformed since the mid 1970s. It was simply useless.

7. Conclusion

We want to end our paper by taking the liberty to voice some more subjective concerns. First, we caution the reader that our own view tends toward skepticism and is arguable. It is not a final word. Researchers with stronger priors about predictability may come away with less skeptical assessments.

Our paper reexamined variables from high-quality papers important enough to have been published in the top academic journals—with manuscript rejection rates as high as 18–19 in 20 papers.[26] Yet, the ongoing performance—given our recycling of the same data in which they were established—does not seem to have been particularly good. As of the mid 2020s, despite our intimate familiarity with these variables and the academic research prediction enterprise, we would not tilt our own investment portfolios with great enthusiasm according to these signals.

We consider it difficult or impossible to assess what historical performance was exceptional and unusual and indicative of future performance. Furthermore, our variables may have performed collectively so poorly, because COVID was an entirely unexpected (positive) shock. (Some variables improved, others deteriorated.) However, analogously, the same caveat applies also to the evidence in the original papers. Perhaps, they performed well just in an original unexpected random realization (like the 1974 bear market), too. Unfortunately, researchers only have one historical realization. There are no reruns.

[26] Welch 2014 suggests that referees and editors do however have a high error rate.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

This raises another question that is difficult or impossible to answer given our short one-time data series: for the variables that are no longer working (perhaps because they involved theories of investor mistakes), did investors read the academic papers and thereby make predictability go away (McLean and Pontiff 2016), or were many of the variables just ex ante spurious to begin with? In principle, one can check whether the IS/OOSCT performance deteriorated *after* the publication of the study. However, we do not know of well-established related statistics in the literature for time-series predictability disappearance. We suspect it would require more data and variables to make a determination, anyway.

We absolutely do not want to imply that the authors of the papers we examined here made inappropriate choices. Indeed, we could replicate the basic performance of all except one paper. Academic research has no better alternative than to evaluate historical data in time. Researchers are not endowed with magic prescience. Our paper enjoyed additional ex post hindsight data.

Instead, we are inclined to agree with Lo and MacKinlay 1990 and Harvey, Liu, and Zhu 2016, who worry more about our collective academic research enterprise, that is, collective "data torture." For every predictive variable stumbled upon and published by a lucky researcher, there are probably hundreds that failed and were never published. They suggest higher null rejection thresholds. Fama's perspective that it is difficult to "beat" very competitive financial markets remains highly appealing *to us*.

The published results in the literature have conveyed a distorted picture of reality, perhaps more obvious to participating producers than to outside consumers. Readers, referees, editors, and journals like papers with impressive results. Who doesn't? In addition to the problem that academic research gives the wrong impression (i.e., that it is easy to predict the stock market), a secondary problem is crowding out. Authors who write more mundane papers, which fail to show remarkable predictive powers, are likely not to be published and thus disappear from the academic rat race. The incentives imposed by our collective on its members are clear. The null rejection thresholds are probably influenced less by a collective assessment of the number of variables that researchers have examined, but more by a mapping of researchers' and referees' priors and the collective (publication) penalty functions themselves.

What about our introductory statement about strong theories? Audiences like impressive results even more so if the results can be justified based on a "strong theoretical basis." Not surprisingly, motivated authors often obligingly offer them. Many of our papers seemed remarkably confident in expressing *strong support* of "theory,"[27] regardless of whether these theories are neoclassical or behavioral. However, the presence of these theories seems not to have offered

[27] For example, papers "point to the importance of theory x," "delayed reaction by investors," "agents recognize marketwide undervaluation," "investors' biased beliefs," "key to identifying predictability," "psychological evidence," and "arguably the strongest predictor."

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

the desired forward-looking performance stability that theory was supposed to provide. Theory itself is pliable.[28] Of course, the alternative of unlimited theoryless freedom is also quite unappealing.

Still, it is our theory that theory has not solved the problem here. Too many theories have been and continue to be the wrong theories. Many of the reexamined papers' published claims defy our (and perhaps common) sense. Many of the models were promising unusually high timing rates of returns to their readers, based on exploiting the ignorance or strange risk preferences of ordinary investors. To us, easy profits forward-looking seems incredible. Large U.S. stocks are traded in highly competitive markets, where even the smartest funds (despite willingness to take on risk on the margin) have struggled to perform well.

The somewhat less aggressive adherents of risk-factor based models—though the nature and the measure of the "risks" in the factors are typically themselves mysteries—have similarly underperformed more often than not. A long procession of academics who have been involved in market-timing and/or stock-selection based funds can attest to this difficulty, too.

We remain comfortable with our original perspective in Goyal and Welch 2008. We are not confident that our data—now updated to late 2021—can tell us which variables will help us predict the equity premium looking forward. This is even though we are risk-neutral investors by at least a modest margin who would be willing to take on more risk.

Appendix

A. Regression Stability and Performance in Halves

Table A.1
First vs. Second Half Stability of IS Coefficients, T-statistics

		Authors'		Our Sample	
Vrbl		First	Second	First	Second
Panel A: Monthly Variables					
BH	vp	3.68^{***}	1.68 ✗	1.32	1.01
BPS	impvar	3.83^{***}	0.66	1.32	1.89
BTZ	vrp	4.29^{***}	2.20^{**}	1.92	−0.05 ✗
CEP	lzrt	1.81^{*}	2.02^{**}	0.70	−0.62 ✗
CP	ogap	-4.32^{***}	−1.33 ✗	−0.24	−3.32
DJM	wtexas	−1.56 ✗	-2.98^{***}	−0.98	−1.09
HJTZ	sntm	1.84^{*}	1.32 ✗	2.11	1.00
JT	ndrbl	-2.52^{**}	−0.99 ✗	−1.81	−0.08
JZZ	skvw	-1.67^{*}	-1.91^{*}	0.56	−1.51 ✗
KJ	tail	1.78^{*}	1.70^{*}	−0.31	1.81 ✗
KP	fbm	2.44^{**}	3.27^{***}	2.96	0.57

[28] Deeper judgment about theories is often dismissed in the vein of Friedman's positive economics that asks theory more to predict outcomes than to make common sense.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Table A.1
Continued

Vrbl		Authors' First	Authors' Second	Our Sample First	Our Sample Second
Panel A: Monthly Variables					
LY	dtoy	1.70*	1.24 ✗	0.59	−0.62 ✗
LY	dtoat	−2.75***	−2.93***	−0.12	−1.11
$Maio_{(13)}$	ygap	1.14 ✗	1.63 ✗	0.41	0.75
$Maio_{(16)}$	rdsp	−0.70 ✗	−2.52**	0.92	−0.12 ✗
Mrtn	rsvix	0.93 ✗	1.98 ✗	−0.63	1.46 ✗
NRTZ	tchi	1.26 ✗	1.40 ✗	1.64	0.72
PW	avgcor	2.02**	1.51 ✗	0.04	1.76
RRZ	shtint	−1.41 ✗	−1.62 ✗	−0.16	−2.04
Y	disag	−2.41**	−4.58***	−1.06	1.36 ✗
BMRR	ntis	n/a	n/a	−1.40	−0.19
CSa	d/p	n/a	n/a	0.84	0.48
CSb	d/y	n/a	n/a	1.07	0.56
CSc	e/p	n/a	n/a	3.07	0.11
CSd	d/e	n/a	n/a	−0.72	0.35 ✗
CSe	svar	n/a	n/a	0.01	−0.42 ✗
Cmpl	tby	n/a	n/a	−1.44	−1.65
FFa	lty	n/a	n/a	−2.16	−1.42
FFb	ltr	n/a	n/a	0.30	1.55
FFc	tms	n/a	n/a	0.54	0.95
FFd	dfy	n/a	n/a	0.05	0.44
FFe	dfr	n/a	n/a	0.08	1.28
FS	infl	n/a	n/a	−0.59	−1.26
KS	b/m	n/a	n/a	1.16	−0.16 ✗

Panel B: Quarterly Variables

Vrbl		Authors' First	Authors' Second	Our Sample First	Our Sample Second
AMP	pce	−2.53**	−2.40**	−2.57	−2.74
BY	govik	2.25**	−0.45 ✗	2.49	−1.85 ✗
CGP	crdstd	−1.71*	−1.44 ✗	−1.98	−0.70
LL	cay	n/a	n/a	2.74	−2.86 ✗
Crn	i/k	n/a	n/a	−3.74	−1.66

Panel C: Annual Variables

Vrbl		Authors' First	Authors' Second	(T [CalY]) First	(T [CalY]) Second	T (JulY) Fir (J)	T (JulY) Sec (J)
CGMS	skew	−1.99**	0.25 ✗	−0.91	2.24 ✗	0.35	1.35
HHT	accrul	1.07 ✗	9.71***	1.08	3.94	1.82	2.73
HHT	cfacc	−2.41**	−1.43 ✗	−3.15	−2.38	−1.38	−1.09
MR	gpce	−3.78***	−2.25**	−2.87	−1.85	−1.20	−2.04
MR	gip	−5.50***	−1.53 ✗	0.11	−1.74 ✗	−0.03	−0.95
PST	house	3.65***	−2.65 ✗	1.03	−0.21 ✗	0.69	−0.19 ✗
BW	eqis	n/a	n/a	−3.61	−0.12	−2.69	0.60 ✗

Explanations: In a stable specification, the two coefficients should be similar. Variables whose predictive coefficient change in sign are followed by a crossmark (✗). The left columns use author sample *and* author specifications, as in Table 2. The right columns use our sample and our specification, as in Table 3. CalY refers to calendar year, JulY to July-to-June years. We did not calculate the original authors' halves for variables used earlier in Baker and Wurgler 2000. This is why they are marked n/a.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

References

Atanasov, V., S. V. Møller, and R. Priestley. 2020. Consumption fluctuations and expected returns. *Journal of Finance* 75:1677–713.

Backus, D., M. Chernov, and S. Zin. 2014. Sources of entropy in representative agent models. *Journal of Finance* 69:51–99.

Baker, M., and J. Wurgler. 2000. The equity share in new issues and aggregate stock returns. *Journal of Finance* 55:2219–57.

———. 2007. Investor sentiment in the stock market. *Journal of Economic Perspectives* 21:129–52.

Bakshi, G., G. Panayotov, and G. Skoulakis. 2011. Improving the predictability of real economic activity and asset returns with forward variances inferred from option portfolios. *Journal of Financial Economics* 100:475–95.

Bekaert, G., and M. Hoerova. 2014. The VIX, the variance premium and stock market volatility. *Journal of Econometrics* 183:181–92.

Belo, Frederico, and J. Yu. 2013. Government investment and the stock market. *Journal of Monetary Economics* 60:325–339.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29:1165–88.

Bollerslev, T., G. Tauchen, and H. Zhou. 2009. Expected stock returns and variance risk premia. *Review of Financial Studies* 22:4463–92.

Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts. 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance* 62:877–915.

Campbell, J. Y. 1987. Stock returns and the term structure. *Journal of Financial Economics* 18:373–99.

Campbell, J. Y., and R. J. Shiller. 1988. Stock prices, earnings, and expected dividends. *Journal of Finance* 43:661–76.

Campbell, J. Y., and S. B. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–31.

Cederburg, S., T. L. Johnson, and M. S. O'Doherty. 2023. On the economic significance of stock return predictability. *Review of Finance* 27:619–57.

Chava, S., M. Gallmeyer, and H. Park. 2015. Credit conditions and stock return predictability. *Journal of Monetary Economics* 74:117–32.

Chen, Y., G. W. Eaton, and B. S. Paye. 2018. Micro (structure) before macro? the predictive power of aggregate illiquidity for stock returns and economic activity. *Journal of Financial Economics* 130:48–73.

Clark, T. E., and M. W. McCracken. 2001. Tests of forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105:85–110.

Cochrane, J. H. 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance* 46:209–37.

Colacito, R., E. Ghysels, J. Meng, and W. Siwasarit. 2016. Skewness in expected macro fundamentals and the predictability of equity returns: Evidence and theory. *Review of Financial Studies* 29:2069–109.

Cooper, I., and R. Priestley. 2009. Time-varying risk premiums and the output gap. *Review of Financial Studies* 22:2801–33.

de Oliveira, T. 2022. Dissecting market expectations in the cross-section of book-to-market ratios. *Critical Finance Review* 11:361–73.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

DeMiguel, V., L. Garlappi, and R. Uppal. 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies* 22:1915–53.

Diebold, F. X., and R. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13:253–63.

Driesprong, G., B. Jacobsen, and B. Maat. 2008. Striking oil: Another puzzle? *Journal of Financial Economics* 89:307–27.

Fama, E. F., and K. R. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25:23–49.

Fama, E. F., and G. W. Schwert. 1977. Asset returns and inflation. *Journal of Financial Economics* 5:115–46.

Giacomini, R., and H. White. 2006. Tests of conditional predictive ability. *Econometrica* 74:1545–78.

Goyal, A., and I. Welch. 2003. Predicting the equity premium with dividend ratios. *Management Science* 49:639–54.

———. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–508.

Guo, H. 2006. On the out-of-sample predictability of stock market returns. *Journal of Business* 79:645–70.

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

Hirshleifer, D., K. Hou, and S. H. Teoh. 2009. Accruals, cash flows, and aggregate stock returns. *Journal of Financial Economics* 91:389–406.

Huang, D., F. Jiang, J. Tu, and G. Zhou. 2015. Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies* 28:791–837.

Jagannathan, R., and Y. Wang. 2007. Lazy investors, discretionary consumption, and the cross-section of stock returns. *Journal of Finance* 62:1623–61.

Jondeau, E., Q. Zhang, and X. Zhu. 2019. Average skewness matters. *Journal of Financial Economics* 134:29–47.

Jones, C. S., and S. Tuzel. 2013. New orders and asset prices. *Review of Financial Studies* 26:115–57.

Kelly, B., and H. Jiang. 2014. Tail risk and asset prices. *Review of Financial Studies* 27:2841–71.

Kelly, B., and S. Pruitt. 2013. Market expectations in the cross-section of present values. *Journal of Finance* 68:1721–56.

———. 2022. Dissecting market expectations in the cross-section of book-to-market ratios: A comment. *Critical Finance Review* 11:375–81.

Kostakis, A., T. Magdalinos, and M. P. Stamatogiannis. 2015. Robust econometric inference for stock return predictability. *Review of Financial Studies* 28:1506–53.

Kothari, S., and J. Shanken. 1997. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics* 44:169–203.

Lettau, M., and S. Ludvigson. 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 56:815–49.

Li, J., and J. Yu. 2012. Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics* 104:401–19.

Lo, A. W. 2002. The statistics of sharpe ratios. *Financial Analysts Journal* 58:36–52.

Lo, A. W., and A. C. MacKinlay. 1990. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3:431–67.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024

Löffler, G. 2022. Equity premium forecasts tend to perform worse against a buy-and-hold benchmark. *Critical Finance Review* 11:65–77.

Maio, P. 2013. The "fed model" and the predictability of stock returns. *Review of Finance* 17:1489–533.

———. 2016. Cross-sectional return dispersion and the equity premium. *Journal of Financial Markets* 29:87–109.

Martin, I. 2011. Simple variance swaps. Working Paper, London School of Economics.

———. 2017. What is the expected return on the market? *Quarterly Journal of Economics* 132:367–433.

Martin, I. W., and S. Nagel. 2022. Market efficiency in the age of big data. *Journal of Financial Economics* 145:154–77.

McCracken, M. W. 2007. Asymptotics for out of sample tests of granger causality. *Journal of Econometrics* 140:719–52.

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

Møller, S. V., and J. Rangvid. 2015. End-of-the-year economic growth and time-varying expected returns. *Journal of Financial Economics* 115:136–54.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou. 2014. Forecasting the equity risk premium: The role of technical indicators. *Management Science* 60:1772–91.

Piazzesi, M., M. Schneider, and S. Tuzel. 2007. Housing, consumption and asset pricing. *Journal of Financial Economics* 83:531–69.

Pollet, J. M., and M. Wilson. 2010. Average correlation and stock market returns. *Journal of Financial Economics* 96:364–80.

Pontiff, J., and L. D. Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of Financial Economics* 49:141–60.

Rapach, D. E., M. C. Ringgenberg, and G. Zhou. 2016. Short interest and aggregate stock returns. *Journal of Financial Economics* 121:46–65.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23:821–62.

Romano, J. P., and M. Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* 73:1237–82.

Seo, S. B., and J. A. Wachter. 2019. Option prices in a model with stochastic disaster risk. *Management Science* 65:3449–69.

Van Binsbergen, J. H., and R. S. Koijen. 2010. Predictive regressions: A present-value approach. *Journal of Finance* 65:1439–71.

Welch, I. 2014. Referee recommendations. *Review of Financial Studies* 27:2773–804.

———. 2016. The (time-varying) importance of disaster risk. *Financial Analysts Journal* 72:14–30.

White, H. 2001. A reality check for data snooping. *Econometrica* 68:1097–126.

Yu, J. 2011. Disagreement and return predictability of stock portfolios. *Journal of Financial Economics* 99:162–83.

Zarnowitz, V., and L. A. Lambros. 1987. Consensus and uncertainty in economic prediction. *Journal of Political Economy* 95:591–621.

Downloaded from https://academic.oup.com/rfs/article/37/11/3490/7749383 by Studienzentrum Himmel+Erde user on 17 October 2024