The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXVIII, NO. 5 • OCTOBER 2023

Is There a Replication Crisis in Finance?

THEIS INGERSLEV JENSEN,* BRYAN KELLY, and LASSE HEJE PEDERSEN

ABSTRACT

Several papers argue that financial economics faces a replication crisis because the majority of studies cannot be replicated or are the result of multiple testing of too many factors. We develop and estimate a Bayesian model of factor replication that leads to different conclusions. The majority of asset pricing factors (i) can be replicated; (ii) can be clustered into 13 themes, the majority of which are significant parts of the tangency portfolio; (iii) work out-of-sample in a new large data set covering 93 countries; and (iv) have evidence that is strengthened (not weakened) by the large number of observed factors.

SEVERAL RESEARCH FIELDS FACE REPLICATION CRISES (or credibility crises), including medicine (Ioannidis (2005)), psychology (Nosek, Spies, and Motyl (2012)), management (Bettis (2012)), experimental economics (Maniadis, Tufano, and List (2017)), and now also financial economics. Challenges to the replicability of finance research take two basic forms:

1. *No internal validity*. Most studies cannot be replicated with the same data (e.g., because of coding errors or faulty statistics) or are not robust in the sense that the main results cannot be replicated using slightly

*Theis Ingerslev Jensen is at Copenhagen Business School. Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER. Lasse Heje Pedersen is at AQR Capital Management, Copenhagen Business School, and CEPR. We are grateful for helpful comments from Nick Barberis; Andrea Frazzini; Cam Harvey (discussant); Antti Ilmanen; Ronen Israel; Andrew Karolyi; John Liew; Toby Moskowitz; Stefan Nagel; Scott Richardson; Anders Rønn-Nielsen; Neil Shephard (discussant); and seminar and conference participants at AFA 2022, NBER 2021, AQR, Georgetown Virtual Fintech Seminar, Tisvildeleje Summer Workshop 2020, Yale, and the CFA Institute European Investment Conference 2020. We thank Tyler Gwinn for excellent research assistance. Jensen and Pedersen gratefully acknowledge support from the Center for Big Data in Finance (grant no. DNRF167). AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. A conflict of interest disclosure statement can be found on *The Journal of Finance* website.

Correspondence: Theis Ingerslev Jensen, Department of Finance, Copenhagen Business School, Solbjerg Plads 3, A4.15, DK-2000 Frederiksberg, Denmark; e-mail: tij.fi@cbs.dk.

DOI: 10.1111/jofi.13249

different methodologies and/or slightly different data.[1] For example, Hou, Xue, and Zhang (2020) state that:
"Most anomalies fail to hold up to currently acceptable standards for empirical finance."

2. *No external validity*. Most studies may be robustly replicated, but are spurious and driven by "*p*-hacking," that is, find significant results by testing multiple hypotheses without controlling the false discovery rate (FDR). Such spurious results are not expected to replicate in other samples or time periods, in part because the sheer number of factors is simply too large, and too fast growing, to be believable. For example, Cochrane (2011) asks for a consolidation of the "factor zoo," and Harvey, Liu, and Zhu (2016) state that:
"most claimed research findings in financial economics are likely false."[2]

In this paper, we examine these two challenges both theoretically and empirically. We conclude that neither criticism is tenable. The majority of factors do replicate, do survive joint modeling of all factors, do hold up out-of-sample, are strengthened (not weakened) by the large number of observed factors, are further strengthened by global evidence, and the number of factors can be understood as multiple versions of a smaller number of themes.

These conclusions rely on new theory and data. First, we show that factors must be understood in light of economic theory, and we develop a Bayesian model that offers a very different interpretation of the evidence on factor replication. Second, we construct a new global data set of 153 factors across 93 countries. To help advance replication in finance, we have made this data set easily accessible to researchers by making our code and data publically available.[3]

Replication results. Figure 1 illustrates our main results and how they relate to the literature in a sequence of steps. It presents the "replication rate," that is, the percent of factors with a statistically significant average excess return.

[1] Hamermesh (2007) contrasts "pure replication" and "scientific replication." Pure replication is "checking on others' published papers using their data," also called "reproduction" by Welch (2019), while scientific replication uses a "different sample, different population and perhaps similar, but not identical model." We focus on scientific replication. We propose a new modeling framework to jointly estimate factor alphas, we use robust factor construction methods that are applied uniformly to all factors, and we test both internal and external validity of prior factor research along several dimensions, including out-of-sample time-series replication and international sample replication. In complementary and contemporaneous work, Chen and Zimmermann (2022) consider pure replication, attempting to use the same data and methods as the original papers for a large number of factors. They are able to reproduce nearly 100% of factors, but Hou, Xue, and Zhang (2020) challenge the scientific replication and Harvey, Liu, and Zhu (2016) challenge validity due to multiple testing (MT).

[2] Similarly, Linnainmaa and Roberts (2018) state that: "the majority of accounting-based return anomalies, including investment, are most likely an artifact of data snooping."

[3] The data are available at https://jkpfactors.com. The data will be updated over time and will also be available through Wharton Research Data Services (WRDS). The code is available at https://github.com/bkelly-lab/ReplicationCrisis.

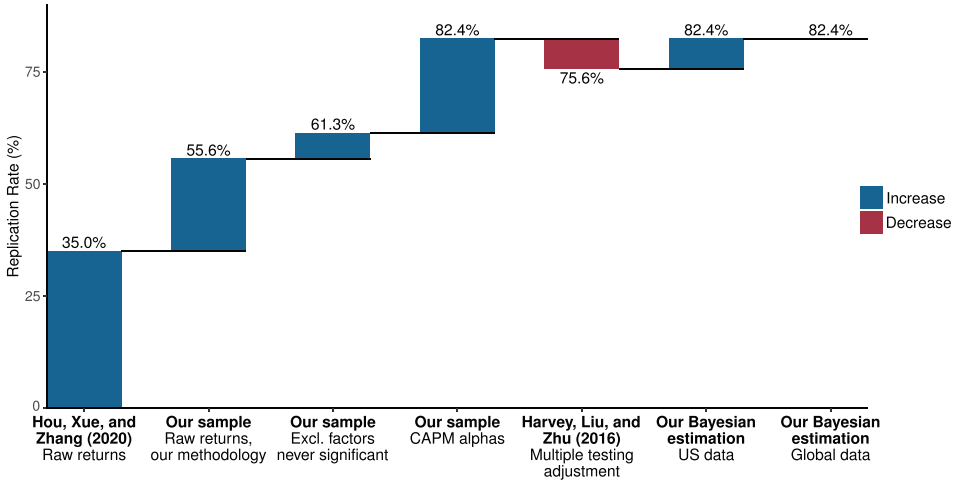

Figure 1. Replication rates versus the literature. This figure summarizes analyses throughout the paper. Refer to Section III for estimation details. (Color figure can be viewed at wileyonlinelibrary.com)

The starting point of Figure 1—the first bar on the left—is the 35% replication rate reported in the expansive factor replication study of Hou, Xue, and Zhang (2020). The second bar in Figure 1 shows a 55.6% baseline replication rate in our main sample of U.S. factors. It is based on significant ordinary least squares (OLS) t-statistics for average raw factor returns, in direct comparability to the 35% calculation from Hou, Xue, and Zhang (2020). This difference arises because our sample is longer, we add 15 factors to our sample that come from prior literature but are not studied by Hou, Xue, and Zhang (2020), and, we believe, minor conservative factor construction details that robustify factor behavior.[4] We discuss this decomposition further in Section II, where we detail our factor construction choices and discuss why we prefer them.

The Hou, Xue, and Zhang (2020) sample includes a number of factors that the original studies find to be insignificant.[5] We exclude these when calculating the replication rate. After making this adjustment, the replication rate rises to 61.3%, as shown in the third bar in Figure 1.

[4] In particular, we use tercile spreads while they use deciles, we use tercile breakpoints from all stocks above the NYSE 20th percentile (i.e., non-micro-caps) while they use straight NYSE breakpoints, we always lag accounting data four months while they use a mixture of updating schemes, we exclude factors based on IBES data due to its relatively short history, we use capped value-weighting while they use straight value-weights, and we look at returns over a one-month holding period while they use one, six, and 12 months. In Section I of the Internet Appendix, we detail how each change affects the replication rate. The Internet Appendix may be found in the online version of this article.

[5] We identify 34 factors from Hou, Xue, and Zhang (2020) for which the original paper did not find a significant alpha or did not study factor returns (see Table IA.II of the Internet Appendix).

Alpha, not raw return. Hou, Xue, and Zhang (2020) analyze and test factors' raw returns, but if we wish to learn about "anomalies," economic theory dictates the use of risk-adjusted returns. The raw return can lead to incorrect inferences for a factor if this return differs from the alpha. When the raw return is significant but the alpha is not, this simply means that the factor is taking risk exposure and the risk premium is significant, which does not indicate anomalous factor returns. Likewise, when the raw return is insignificant but the alpha is significant, the factor's efficacy is masked by its risk exposure. An example of this is the low-beta anomaly, whereby theory predicts that the alpha of a dollar-neutral low-beta factor is positive but its raw return is negative or close to zero (Frazzini and Pedersen (2014)). In this case, the "failure to replicate" of Hou, Xue, and Zhang (2020) actually supports the betting-against-beta theory. We analyze the alpha to the capital asset pricing model (CAPM), which is the clearest theoretical benchmark model that is not mechanically linked to other so-called anomalies in the list of replicated factors. The fourth bar in Figure 1 shows that the replication rate rises to 82.4% based on tests of factors' CAPM alpha.

MT and our Bayesian model. The first four bars in Figure 1 are based on individual OLS t-statistics for each factor. But Harvey, Liu, and Zhu (2016) rightly point out that this type of analysis suffers from an MT problem. Harvey, Liu, and Zhu (2016) recommend MT adjustments that raise the threshold for a t-statistic to be considered statistically significant. We report one such MT correction using a leading method proposed by Benjamini and Yekutieli (2001). Accounting for MT in this manner, we find that the replication rate drops to 75.6% (the fifth bar of Figure 1). For comparison, Hou, Xue, and Zhang (2020) consider a similar adjustment and find that their replication rate drops from 35% with OLS to 18% after MT correction.

However, common frequentist MT corrections can be unnecessarily crude. Our handling of the MT problem is different. We propose a Bayesian framework for the joint behavior of all factors, resulting in an MT correction that sacrifices much less power than its frequentist counterpart, which we demonstrate via simulation.[6] To understand the benefits of our approach, note first that we impose a prior that all alphas are expected to be zero. The role of the Bayesian prior is conceptually similar to that of frequentist MT corrections—it imposes conservatism on statistical inference and controls the FDR. Second, our *joint* factor model allows us to conduct inference for all factor alphas simultaneously. The joint structure among factors leverages dependence in the data to draw more informative statistical inferences (relative to conducting independent individual tests). Our zero-alpha prior shrinks alpha estimates

[6] A large statistics literature (see Gelman et al. (2013) and references therein) explains how Bayesian estimation naturally addresses MT problems and Gelman, Hill, and Yajima (2012) conclude that "the problem of multiple comparisons can disappear entirely when viewed from a hierarchical Bayesian perspective." Chinco, Neuhierl, and Weber (2021) use a Bayesian estimation framework similar to ours for a different (but conceptually related) problem. They infer the distribution of coefficients in a stock return prediction model to calculate what they refer to as the "anomaly base rate."

of all factors, leading to fewer discoveries (i.e., a lower replication rate), with similar conservatism as a frequentist MT correction. At the same time, however, the model allows us to learn more about the alpha of any individual factor, borrowing estimation strength across all factors, and the improved precision of alpha estimates for all factors can increase the number of discoveries. Which effect dominates when we construct our final Bayesian model—the conservative shrinkage to the prior or the improved precision of alphas—is an empirical question.

In our sample, we find that the two effects exactly offset on average, which is why the Bayesian MT view delivers a replication rate identical to the OLS-based rate. Specifically, our estimated replication rate rises to 82.4% (the sixth bar of Figure 1) using our Bayesian approach to the MT problem.[7] The intuition behind this surprising result is simply that having many factors (a "factor zoo") can be a strength rather than a weakness when assessing the replicability of factor research. It is obvious that our posterior is tighter when a factor has performed better and has a longer time series. But the posterior is further tightened if similar factors have also performed well and if additional data show that these factors have performed well in many other countries.[8]

Benefits of our model beyond the replication rate. One of the key benefits of Bayesian statistics is that one recovers not just a point estimate but rather the entire posterior distribution of parameters. The posterior allows us to make any possible probability calculation about parameters. For example, in addition to the replication rate, we calculate the posterior probability of false discoveries (FDR) and the posterior expected fraction of true factors. Moreover, we calculate Bayesian confidence intervals (also called credibility intervals) for each of these estimates. We find that our 82.4% replication rate has a tight posterior standard error of 2.8%. The posterior Bayesian FDR is only 0.1% with a 95% confidence interval of [0.0%, 1.0%], demonstrating the small risk of false discoveries. The expected fraction of true factors is 94.0% with a posterior standard error of 1.3%.

Global replication. Having found a high degree of internal validity of prior research, we next consider external validity across countries and over time. Regarding the former, we investigate how our conclusions are affected when we extend the data to include all factors in a large global panel of 93 countries. The last bar in Figure 1 shows that based on the global sample, the final replication rate is 82.4%. This estimate is based on the Bayesian model applied to a sample of global factors that weights country-specific factors in proportion to the country's total market capitalization. The model continues to account for MT. The global result shows that factor performance in the United States

[7] Our Bayesian approach leads to an even larger increase in the replication rate when using pure value-weighted returns (see Figure IA.1 of the Internet Appendix) and when considering global evidence outside the United States (as we show later, in Figure 6).

[8] Taking this intuition further, we can obtain additional information from studying whether factors work in other asset classes, as has been done for value and momentum (Asness, Moskowitz, and Pedersen (2013)), betting against beta (Frazzini and Pedersen (2014)), time-series momentum (Moskowitz, Ooi, and Pedersen (2012)), and carry (Koijen et al. (2018)).

replicates well in an extensive cross section of countries. Serving as our final estimate, the global factor replication rate more than doubles that of Hou, Xue, and Zhang (2020) by grounding our tests in economic theory and modern Bayesian statistics. We conclude from the global analysis that factor research demonstrates external validity in the cross section of countries.

Postpublication performance. McLean and Pontiff (2016) find that U.S. factor returns "are 26% lower out-of-sample and 58% lower post-publication."[9] Our Bayesian framework shows that, given a prior belief of zero alpha but an OLS alpha ($\hat{\alpha}$) that is positive, our posterior belief about alpha lies somewhere between zero and $\hat{\alpha}$. Hence, a positive but attenuated postpublication alpha is the expected outcome based on Bayesian learning, rather than a sign of nonreproducibility. Further, when comparing factors cross-sectionally, the prediction of the Bayesian framework is that higher prepublication alphas, if real, should be associated with higher postpublication alphas on average. This is what we find. We present new and significant cross-sectional evidence that factors with higher in-sample alpha generally have higher out-of-sample alpha. The attenuation in the data is somewhat stronger than predicted by our Bayesian model. We conclude that factor research demonstrates external validity in the time series, but there appears to be some decay of the strongest factors that could be due to arbitrage or data mining.[10]

Publication bias. We also address the issue that factors with strong in-sample performance are more likely to be published while poorly performing factors are more likely to be unobserved in the literature. Publication bias can influence our full-sample Bayesian evidence through the empirical Bayes (EB) estimation of prior hyperparameters. To account for this bias, we show how to pick a prior distribution that is unaffected by publication bias by using only out-of-sample data or estimates from Harvey, Liu, and Zhu (2016). Using such priors, the full-sample alphas are shrunk more heavily toward zero. The result is a slight drop in the U.S. replication rate to 81.5%. If we add an extra degree of conservatism to the prior, the replication rate drops to 79.8%. Further, our out-of-sample evidence over time and across countries is not subject to publication bias.

Multidimensional challenge: A Darwinian view of the factor zoo. Harvey, Liu, and Zhu (2016) challenge the sheer number of factors, which Cochrane (2011) refers to as "the multidimensional challenge." We argue that the factor universe should not be viewed as hundreds of distinct factors. Instead, factors cluster into a relatively small number of highly correlated themes. This property features prominently in our Bayesian modeling approach. Specifically, we propose a factor taxonomy that algorithmically classifies factors into 13 themes

[9] Extending the evidence to global stock markets, Jacobs and Müller (2020) find that "the United States is the only country with a reliable post-publication decline in long-short returns." Chen and Zimmermann (2020b) use Bayesian methods to estimate bias-corrected postpublication performance and find that average returns drop by only 12% after publication in U.S. data.

[10] Data prior to the sample used in original studies also constitute out-of-sample evidence (Linnainmaa and Roberts (2018), Ilmanen et al. (2021)). Our external validity conclusions hold when we also include pre–original-study out-of-sample evidence.

possessing a high degree of within-theme return correlation and economic concept similarity, and low across-theme correlation. The emergence of themes in which factors are minor variations on a related idea is intuitive. For example, each value factor is defined by a specific valuation ratio, but there are many plausible ratios. Considering their variations is not spurious alpha-hacking, particularly when the "correct" value signal construction is debatable.

We estimate a replication rate greater than 50% in 11 of the 13 themes (based on the Bayesian model including MT adjustment), the exceptions being "low leverage" and "size" factor themes. We also analyze which themes matter when simultaneously controlling for all other themes. To do so, we estimate the ex post tangency portfolio of 13 theme-representative portfolios. We find that 10 of the 13 themes enter into the tangency portfolio with significantly positive weights, where the three displaced themes are "profitability," "investment," and "size."

Why, the profession asks, have we arrived at a "factor zoo"?[11] Evidently the answer is because the risk-return trade-off is complex and difficult to measure. The complexity manifests in our inability to isolate a single silver-bullet characteristic that pins down the risk-return trade-off. Classifying factors into themes, we trace the economic culprits to roughly a dozen concepts. This is already a multidimensional challenge, but it is compounded by the fact that within a theme there are many detailed choices for how to configure the economic concept, which results in highly correlated within-theme factors. Together, the themes (and the factors in them) each make slightly different contributions to our collective understanding of markets. A more positive take on the factor zoo is *not* as a collective exercise in data mining and false discovery, but rather as a natural outcome of a decentralized effort in which researchers make contributions that are correlated with, but incrementally improve on, the body of knowledge.

Economic implications. Our findings have broad implications for finance researchers and practitioners. We confirm that the body of finance research contains a multitude of replicable information about the determinants of expected returns. Further, we show that investors would have profited from factors deemed significant by our Bayesian method but insignificant by the frequentist MT method proposed by Harvey, Liu, and Zhu (2016). Figure 2 plots the out-of-sample returns of the subset of factors discovered by our method but discarded by the frequentist method. As can be seen, these factors produce an annualized information ratio (IR) of 0.93 in the Unites States and 1.10 globally (ex-U.S.) over the full sample, with t-statistics above five. If we restrict analysis to the sample after that of Harvey, Liu, and Zhu (2016), the performance

[11] See Bryzgalova, Huang, and Julliard (2023), Kelly, Pruitt, and Su (2019), Chordia, Goyal, and Saretto (2020), Kozak, Nagel, and Santosh (2020), Green, Hand, and Zhang (2017), and Feng, Giglio, and Xiu (2020) for other perspectives on high-dimensional asset pricing problems, and Chen (2021) for an argument regarding why p-hacking cannot explain the existence of so many significant factors.

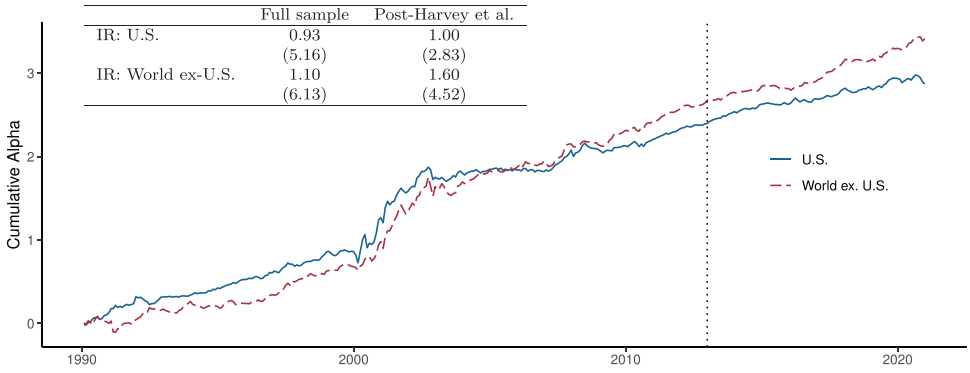

Figure 2. Out-of-sample performance of marginally significant factors. The figure shows the cumulative CAPM alpha of an average of factors significant under our empirical Bayes framework, but not with the Benjamini-Yekutieli adjustment suggested by Harvey, Liu, and Zhu (2016). The significance cutoffs are reestimated each year with the available data. Factors are eligible for inclusion after the sample period in the original paper, so all returns are out-of-sample. The table shows the IR (alpha divided by residual volatility) for the full sample (1990 to 2020) and the post–Harvey, Liu, and Zhu (2016) sample (2013 to 2020) with *t*-statistics in parentheses. The vertical dotted line is at December 2012. (Color figure can be viewed at wileyonlinelibrary.com)

differential remains large and significant.[12] These findings show strong external validity (postoriginal publications, post–Harvey, Liu, and Zhu (2016), different countries) and significant economic benefits of exploiting the joint information in all factor returns rather than simply applying a high cutoff for *t*-statistics. We also show that the optimal risk-return profile has improved over time as factors have been discovered. In other words, the Sharpe ratio of the tangency portfolio has meaningfully increased over time as truly novel determinants of returns have been discovered. These findings can help inform asset pricing theory.

The paper proceeds as follows. Section I describes our Bayesian model of factor replication. Section II presents our new public data set of global factors. Section III contains our empirical assessment of factor replicability. Section IV concludes.

I. A Bayesian Model of Factor Replication

This section presents our Bayesian model for assessing factor replicability. We first draw out some basic implications of the Bayesian framework for interpreting evidence on individual factor alphas. We then present a hierarchical structure for simultaneously modeling factors in a variety themes and across many countries.

[12] The out-of-sample performance across all significant factors under EB is also highly significant as shown in Figure IA.2 of the Internet Appendix.

A. Learning about Alpha: The Bayes Case

A.1. Posterior Alpha

We begin by considering an excess return factor, f_t. A study of "anomalous" factor returns requires a risk benchmark, without which we cannot separate distinctive factor behavior from run-of-the-mill risk compensation. We assume a CAPM benchmark due to its history as a factor research benchmark for decades, and because it is not mechanically related to any of the factors that we attempt to replicate (in contrast to, say, the model of Fama and French (1993), which by construction explains size and value factors). A factor's net performance versus the excess market factor (r_t^m) is its α,

$$f_t = \alpha + \beta r_t^m + \varepsilon_t. \tag{1}$$

Our Bayesian prior is that the alpha is normally distributed with mean zero and variance τ^2, or $\alpha \sim N(0, \tau^2)$. The mean of zero implies that CAPM holds on average, and τ governs potential deviations from CAPM. Intuitively, the higher the confidence in the prior, the lower is τ. The error term, $\varepsilon_t \sim N(0, \sigma^2)$, has volatility σ and is independent and identically distributed over time, and σ and β are observable.[13]

The risk-adjusted return, α, is estimated as the average market-adjusted factor return from T periods of data,

$$\hat{\alpha} = \frac{1}{T}\sum_t \left(f_t - \beta r_t^m\right) = \alpha + \frac{1}{T}\sum_t \varepsilon_t. \tag{2}$$

This observed OLS estimate $\hat{\alpha}$ is distributed $N(\alpha, \sigma^2/T)$ given the true alpha, α. From Bayes' rule, we can compute the posterior distribution of the true alpha given the empirical evidence and prior. The posterior exhaustively describes the Bayesian's beliefs about alpha at a future time $t > T$ given past experience, including the posterior expected factor performance,

$$E(\alpha|\hat{\alpha}) = E\left(f_t - \beta r_t^m \middle| \hat{\alpha}\right). \tag{3}$$

We derive the posterior alpha distribution via Bayes' rule (the derivation, which is standard, is shown in Appendix A). The posterior alpha is normal with mean

$$E(\alpha|\hat{\alpha}) = \kappa\hat{\alpha}, \tag{4}$$

[13] Here, we seek to derive some simple expressions that illustrate the economic implications of Bayesian logic. In the empirical implementation, we use a slightly richer model, as discussed further below. The empirical implementation normalizes factors so that σ is given at 10% for all factors, while β must be estimated, but this does not affect the economic points that we make in this section.

where κ is a shrinkage factor given by

$$\kappa = \frac{\tau^2}{\tau^2 + \sigma^2/T} = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T}} \in (0, 1) \tag{5}$$

and the posterior variance is

$$\text{Var}(\alpha|\hat{\alpha}) = \kappa \frac{\sigma^2}{T} = \frac{1}{\frac{1}{\sigma^2/T} + \frac{1}{\tau^2}}. \tag{6}$$

The first insight from this posterior is that a Bayesian predicts that future returns will have smaller alpha (in absolute value) than the OLS estimate $\hat{\alpha}$, because the posterior mean ($\kappa\hat{\alpha}$) must lie between $\hat{\alpha}$ and the prior mean of zero. Put differently, a large observed alpha might be due to luck. Given the prior, we expect that at least part of this performance is indeed luck. The more data we have (higher T), the less shrinkage there is (i.e., κ closer to one), while the stronger is the prior of zero alpha (i.e., lower τ), the heavier is the shrinkage. We can think of the prior τ in terms of the number of time periods of evidence that it corresponds to. That is, the posterior mean, $E(\alpha|\hat{\alpha})$, corresponds to first observing σ^2/τ^2 time periods with an average alpha of zero, followed by T time periods with a average alpha of $\hat{\alpha}$.

When evaluating out-of-sample evidence, a positive but lower alpha is sometimes interpreted as a sign of replication failure. But this is the expected outcome from the Bayesian perspective (i.e., based on the latest posterior) and can be fully consistent with a high degree of replicability. In fact, postpublication results (as also studied by McLean and Pontiff (2016)) have tended to confirm the Bayesian's beliefs and as a result the Bayesian posterior alpha estimate has been extraordinarily stable over time (see Section III.B.2).

A.2. Alpha-Hacking

Because out-of-sample alpha attenuation is not generally a sign of replication failure, we may want a more direct probe for nonreplicability. We can build such a test into our Bayesian framework by embedding scope for "alpha-hacking," or selectively reporting or manipulating data to artificially make the alpha seem larger. We represent this idea using the following distribution of factor returns in the in-sample period $t = 1, \ldots, T$:

$$f_t = \alpha + \beta r_t^m + \underbrace{\tilde{\varepsilon}_t + u}_{\varepsilon_t}. \tag{7}$$

Here, $\tilde{\varepsilon}_t \sim N(0, \sigma^2)$ captures usual return shocks and $u \sim N(\bar{\varepsilon}, \sigma_u^2)$ represents return inflation due to alpha-hacking. The total in-sample return shock ε_t is normally distributed, $N(\bar{\varepsilon}, \bar{\sigma}^2)$, where $\bar{\varepsilon} \geq 0$ is the alpha-hacking bias, and the variance $\bar{\sigma}^2 = \sigma^2 + \sigma_u^2 \geq \sigma^2$ is elevated due to the artificial noise created by

alpha-hacking.[14] Naturally, the false benefits of alpha-hacking disappear in out-of-sample data, or in other words, $\varepsilon_t \sim N(0, \sigma^2)$ for $t > T$. The Bayesian accounts for alpha-hacking as follows.

PROPOSITION 1 (Alpha-Hacking): *The posterior alpha with alpha-hacking is given by*

$$E(\alpha|\hat{\alpha}) = -\kappa_0 + \kappa^{hacking}\hat{\alpha}, \tag{8}$$

where $\kappa^{hacking} = \frac{1}{1+\frac{\bar{\sigma}^2}{\tau^2 T}} \leq \kappa$ *and* $\kappa_0 = \kappa^{hacking}\bar{\varepsilon} \geq 0$. *Further,* $\kappa^{hacking} \to 0$ *in the limit of "pure alpha-hacking,"* $\tau \to 0$ *or* $\bar{\sigma} \to \infty$.

The Bayesian posterior alpha accounts for alpha-hacking in two ways. First, the estimated alpha is shrunk more heavily toward zero since the factor κ^{hacking} is now smaller. Second, the alpha is further discounted by the intercept term κ_0 due to the bias in the error terms.

We examine alpha-hacking empirically in Section III.B in light of Proposition 1. We consider a cross-sectional regression of factors' out-of-sample (e.g., postpublication) alphas on their in-sample alphas, looking for the signatures of alpha-hacking in the form of a negative intercept term or a slope coefficient that is too small. In addition, Section III.C.2 shows how to estimate the Bayesian model in a way that is less susceptible to the effects of alpha-hacking. Appendix A presents additional theoretical results characterizing alpha-hacking.

B. Hierarchical Bayesian Model

B.1. Shared Alphas: The Case of Complete Pooling

We now embed a critical aspect of factor research into our Bayesian framework: Factors are often correlated and conceptually related to each other. For concreteness, we begin with a setting in which the researcher has access to "domestic" evidence in (1) as well as "global" evidence from an international factor, f_t^g, with known exposure β^g to the global market index r_t^g:

$$f_t^g = \alpha + \beta^g r_t^g + \varepsilon_t^g. \tag{9}$$

Here, we assume that the true alpha for this global factor is the same as the domestic alpha. In other words, we have complete "pooling" of information about alpha across the two samples. As an alternative interpretation, the researcher could have access to two related factors, say, two different value factors in the same country, and assume that they have the same alpha because they capture the same investment principle.

[14] We note that this elevated variance cannot be detected by looking at the in-sample variance of residual returns since the alpha-hacking term u does not depend on time t.

The global shock, ε_t^g, is normally distributed $N(0, \sigma^2)$, and ε_t^g and ε_t are jointly normal with correlation ρ.[15] The estimated alpha based on the global evidence is simply its market-adjusted return:

$$\hat{\alpha}^g = \frac{1}{T}\sum_t \left(f_t^g - \beta^g r_t^g\right). \tag{10}$$

To see the power of global evidence (or, more generally, the power of observing related strategies), we consider the posterior when observing both the domestic and global evidence.

PROPOSITION 2 (The Power of Shared Evidence): *The posterior alpha given the domestic estimate, $\hat{\alpha}$, and the global estimate, $\hat{\alpha}^g$, is normally distributed with mean*

$$E(\alpha|\hat{\alpha}, \hat{\alpha}^g) = \kappa^g\left(\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g\right). \tag{11}$$

The global shrinkage parameter is

$$\kappa^g = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T}\frac{1+\rho}{2}} \in [\kappa, 1], \tag{12}$$

which decreases with the correlation ρ, attaining the minimum value, $\kappa^g = \kappa$, when $\rho = 1$. The posterior variance is lower when observing both domestic and global evidence:

$$\text{Var}(\alpha|\hat{\alpha}) \geq \text{Var}(\alpha|\hat{\alpha}, \hat{\alpha}^g). \tag{13}$$

Naturally, the posterior depends on the average alpha observed domestically and globally. Furthermore, the combined alpha is shrunk toward the prior of zero. The shrinkage factor κ^g is smaller (heavier shrinkage) if the markets are more correlated because the global evidence provides less new information. With low correlation, the global evidence adds a lot of independent information, shrinkage is lighter, and the Bayesian becomes more confident in the data and less reliant on the prior. The proposition shows that if a factor has been found to work both domestically and globally, then the Bayesian expects stronger out-of-sample performance than a factor that has only worked domestically (or has only been analyzed domestically).

Two important effects are at play here, and both are important for understanding the empirical evidence presented below: The domestic and global alphas are shrunk both toward *each other* and toward *zero*. For example, suppose that a factor worked domestically but not globally, say, $\hat{\alpha} = 10\% > \hat{\alpha}^g = 0\%$.

[15] The framework can be generalized to a situation in which the global shocks have a different volatility and sample length. In this case, the Bayesian posterior puts more weight on the sample with lower volatility and longer length.

Then, the overall evidence points to an alpha of $\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g = 5\%$ but shrinkage toward the prior results in a lower posterior, say, 2.5%. Hence, the Bayesian expects future factor returns in both regions of 2.5%. The fact that shared alphas are shrunk together is a key feature of a *joint* model, and it generally leads to different conclusions than when factors are evaluated independently. We next consider a perhaps more realistic model in which factors are only partially shrunk toward each other.

B.2. Hierarchical Alphas: The Case of Partial Pooling

We now consider several factors, numbered $i = 1, \ldots, N$. Factor i has a true alpha given by

$$\alpha^i = c + w^i. \tag{14}$$

Here, c is the common component of all alphas, which has a prior distribution given by $N(0, \tau_c^2)$. Likewise, w^i is the idiosyncratic alpha component, which has a prior distribution given by $N(0, \tau_w^2)$, independent of c and across i. Put differently, we can imagine that nature first picks the overall c from $N(0, \tau_c^2)$ and then picks the factor-specific α^i from $N(c, \tau_w^2)$.

This hierarchical model is a realistic compromise between assuming that all factor alphas are completely different (using equation (4) for each alpha separately) and assuming that they are all the same (using Proposition 2). Rather than assuming no pooling or complete pooling, the hierarchical model allows factors to have a common component and an idiosyncratic component.

Suppose we observe factor returns of

$$f_t^i = \alpha^i + \beta^i r_t^m + \varepsilon_t^i, \tag{15}$$

where ε_t^i are normally distributed with mean zero and variance σ^2, and $\text{Cor}(\varepsilon_t^i, \varepsilon_t^j) = \rho \geq 0$ for all i, j.[16] Computing the observed alpha estimates as above, $\hat{\alpha}^i = \frac{1}{T}\sum_t (f_t^i - \beta^i r_t^m)$, we derive the posterior in the following result.[17]

PROPOSITION 3 (Hierarchical Alphas): *The posterior alpha of factor i given the evidence on all factors is normally distributed with mean*

$$E(\alpha^i | \hat{\alpha}^1, \ldots, \hat{\alpha}^N) = \frac{1}{1 + \frac{\rho\sigma^2}{\tau_c^2 T} + \frac{\tau_w^2 + (1-\rho)\sigma^2/T}{\tau_c^2 N}}\,\hat{\alpha}^{\cdot} + \frac{1}{1 + \frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \frac{1}{1 + \frac{\tau_w^2 + (1-\rho)\sigma^2/T}{(\tau_c^2 + \rho\sigma^2/T)N}}\,\hat{\alpha}^{\cdot}\right), \tag{16}$$

[16] Alternatively, we can write the error terms in a similar way as we write the alphas in (14), namely, $\varepsilon_t^i = \sqrt{\rho}\,\bar{\varepsilon}_t + \sqrt{1-\rho}\,\tilde{\varepsilon}_t^i$, where $\tilde{\varepsilon}_t^i$ are idiosyncratic shocks that are independent across factors and of the common shock $\bar{\varepsilon}_t$, with $\text{Var}(\tilde{\varepsilon}_t^i) = \text{Var}(\bar{\varepsilon}_t) = \sigma^2$. Note that we require (the empirically realistic case) that $\rho \geq 0$ since we cannot have an arbitrarily large number of normal random variables with equal negative correlation (because the corresponding variance-covariance matrix would not be positive semidefinite for large enough N).

[17] The general hierarchical model is used extensively in the statistics literature (see, e.g., Gelman et al. (2013)), but to our knowledge the results in Proposition 3 are not in the literature.

where $\hat{\alpha}^{\cdot} = \frac{1}{N}\sum_j \hat{\alpha}^j$ *is average alpha. When the number of factors N grows, the limit is*

$$\lim_{N\to\infty} E(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) = \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}}\,\hat{\alpha}^{\cdot} + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \hat{\alpha}^{\cdot}\right). \tag{17}$$

The posterior variance of factor i's alpha using the information in all factor returns is lower than the posterior variance when looking at this factor in isolation,

$$\mathrm{Var}(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) < \mathrm{Var}(\alpha^i|\hat{\alpha}^i). \tag{18}$$

The posterior variance is decreasing in N and, as $N\to\infty$*, its limit is*

$$\mathrm{Var}(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) \searrow \frac{\rho\sigma^2}{T}\frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}} + \frac{(1-\rho)\sigma^2}{T}\frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}. \tag{19}$$

The main insight of this proposition is that having data on many factors is helpful for estimating the alpha of any of them. Intuitively, the posterior for any individual alpha depends on all of the other observed alphas because they are all informative about the common alpha component. Put differently, the other observed alphas tell us whether alpha exists in general, that is, whether the CAPM appears to be violated in general. Further, the factor's own observed alpha tells us whether this specific factor appears to be especially good or bad. Using all of the factors jointly reduces posterior variance for all alphas. In summary, the joint model with hierarchical alphas has the dual benefits of identifying the common component in alphas and tightening confidence intervals by sharing information among factors.

To understand the proposition in more detail, consider first the (unrealistic) case in which all factor returns have independent shocks ($\rho = 0$). In this case, we essentially know the overall alpha when we see many uncorrelated factors. Indeed, the average observed alpha becomes a precise estimator of the overall alpha with more and more observed factors, $\hat{\alpha}^{\cdot} \to c$. Since we essentially know the overall alpha in this limit, the first term in (17) becomes $1\times\hat{\alpha}^{\cdot}$ when $\rho = 0$, meaning that we do not need any shrinkage here. The second term is the outperformance of factor i above the average alpha, and this outperformance is shrunk toward our prior of zero. Indeed, the outperformance is multiplied by a number less than one, and this multiplier naturally decreases in the return volatility σ and in our conviction in the prior (increases in τ_w).

The posterior variance is also intuitive in the case $\rho = 0$. The posterior variance is clearly lower compared to only observing the performance of factor i itself,

$$\mathrm{Var}(\alpha^i|\hat{\alpha}^1,\hat{\alpha}^2\ldots) = \frac{\sigma^2}{T}\frac{1}{1+\frac{\sigma^2}{\tau_w^2 T}} < \frac{\sigma^2}{T}\frac{1}{1+\frac{\sigma^2}{(\tau_c^2+\tau_w^2)T}} = \mathrm{Var}(\alpha^i|\hat{\alpha}^i), \tag{20}$$

based on (19) and (6). With partial pooling, the posterior variance decreases because the denominator on the left does not have τ_c^2, reflecting that uncertainty about the general alpha has been eliminated by observing many factors.

In the realistic case in which factor returns are correlated ($\rho > 0$), we see that both the average alpha $\hat{\alpha}^{\cdot}$ and factor i's outperformance $\hat{\alpha}^i - \hat{\alpha}^{\cdot}$ are shrunk toward the prior of zero. This is because we cannot precisely estimate the overall alpha even with an infinite number of correlated factors—the correlated part never vanishes. Nevertheless, we still shrink the confidence interval, $\text{Var}(\alpha^i|\hat{\alpha}^1, \ldots, \hat{\alpha}^N) \leq \text{Var}(\alpha^i|\hat{\alpha}^i)$, since more information is always better than less.

B.3. Multilevel Hierarchical Model

The model development to this point is simplified to draw out its intuition. Our empirical implementation is based on a more realistic (and slightly more complex) model that accounts for the fact that factors naturally belong to different economic themes and to different regions.

In our global analysis, we have N different characteristic signals (e.g., book-to-market) across K regions, for a total of NK factors (e.g., U.S., developed, and emerging markets versions of book-to-market). Each of the N signals belongs to a smaller number of J theme clusters, where one cluster consists of various value factors, another consists of various momentum factors, and so on. One level of our hierarchical model allows for partially shared alphas among factors in the same theme cluster. Another level allows for commonality across regions among factors associated with the same underlying characteristic, capturing, for example, the connections between the book-to-market factors in different markets.

Mathematically, this means that an individual factor i has an alpha of

$$\alpha^i = \alpha^o + c^j + s^n + w^i. \tag{21}$$

To illustrate, suppose factor $i \in \{1, \ldots, NK\}$ is the book-to-market factor in the U.S. region. Part of its alpha is driven by a component that is common to all factors, α^o, which we dogmatically fix at zero to be conservative. In addition, this factor i belongs to the value cluster $j \in \{1, \ldots, J\}$, which contributes a cluster-specific alpha $c^j \sim N(0, \tau_c^2)$. Next, since factor i is based on book-to-market characteristic $n \in \{1, \ldots, N\}$, it has an incremental signal-specific alpha of $s^n \sim N(0, \tau_s^2)$ that is shared across regions—for example, it is the common behavior among book-to-market factors regardless of geography. Finally, $w^i \sim N(0, \tau_w^2)$ is factor i's idiosyncratic alpha, namely, the incremental alpha that is unique to the U.S. version of book-to-market.

We write this model in vector form as[18]

$$\alpha = \alpha^o \, 1_{NK} + Mc + Zs + w, \tag{22}$$

[18] The notation 1_N refers to an $N \times 1$ vector of ones and I_N is the $N \times N$ identity matrix.

where $\alpha = (\alpha^1, \ldots, \alpha^{NK})'$, $c = (c^1, \ldots, c^J)'$, $s = (s^1, \ldots, s^N)'$, $w = (w^1, \ldots, w^{NK})'$, M is the $NK \times J$ matrix of cluster memberships, and Z is the $NK \times N$ matrix indicating the characteristic that factor i is based on. In particular, $M_{i,j} = 1$ if factor i is in cluster j and $M_{i,j} = 0$ otherwise. Likewise, $Z_{i,n} = 1$ if factor i is based on characteristic n and $Z_{i,n} = 0$ otherwise. This hierarchical model implies that the prior variance of alpha, denoted by Ω, is[19]

$$\Omega = \text{Var}(\alpha) = MM'\tau_c^2 + ZZ'\tau_s^2 + I_{NK}\tau_w^2. \tag{23}$$

In some cases, we analyze this model within a single region, $K = 1$ (e.g., in our U.S.-only analysis). In this case, there is no difference between signal-specific alphas and idiosyncratic alphas, so we collapse one level of the model by setting $\tau^s = 0$ and $s^n = 0$ for $n \in \{1, \ldots, N\}$. In any case, the following result shows how to compute the posterior distribution of all alphas based on the prior uncertainty, Ω, and a general variance-covariance matrix of return shocks, $\Sigma = \text{Var}(\varepsilon)$. This result is at the heart of our empirical analysis.

PROPOSITION 4: *In the multilevel hierarchical model, the posterior of the vector of true alphas is normally distributed with posterior mean*

$$E(\alpha|\hat{\alpha}) = \left(\Omega^{-1} + T\Sigma^{-1}\right)^{-1}\left(\Omega^{-1}1_{NK}\alpha_0 + T\Sigma^{-1}\hat{\alpha}\right) \tag{24}$$

and posterior variance

$$Var(\alpha|\hat{\alpha}) = \left(\Omega^{-1} + T\Sigma^{-1}\right)^{-1}. \tag{25}$$

As noted above, we set the mean prior alpha to zero ($\alpha_0 = 0$) in our empirical implementation. This prior is based on economic theory and leads to a conservative shrinkage toward zero as seen in (24). We note that, in the data, the observed alphas are mostly positive, not centered around zero. However, these positive alphas are related to the way that factors are signed, that is, according to the convention in the original paper, which almost always leads to a positive factor return in the original sample. If we view this signing convention as somewhat arbitrary, then a symmetry argument implies that a prior of zero is again natural. Put differently, factor means would be centered around zero if we changed signs arbitrarily, so our prior is agnostic about these signs.

C. Bayesian Multiple Testing and Empirical Bayes Estimation

Frequentist MT corrections embody a principle of conservatism that seeks to limit false discoveries by controlling the family-wise error rate (FWER) or

[19] Stated differently, each diagonal element of Ω is $\tau_c^2 + \tau_s^2 + \tau_w^2$. Further, if $i \neq k$, then the $(i,k)^{\text{th}}$ element of Ω is $\tau_c^2 + \tau_s^2$ if i and k are constructed from the same signal in the same cluster in different regions, it is τ_c^2 if i and k are constructed from different signals in the same cluster, and it is zero if i and k are in different clusters.

the FDR. Leading frequentist methods do so by widening confidence intervals and raising *p*-values, but do not alter the underlying point estimate.

C.1. Bayesian Multiple Testing

A large statistics literature shows that Bayesian modeling is effective for making reliable inferences in the face of MT.[20] Drawing on this literature, our hierarchical model is a prime example of how Bayesian methods accomplish their MT correction based on two key model features.

The first such feature is the model prior, which imposes statistical conservatism in analogy to frequentist MT methods. It anchors the researcher's beliefs to a sensible default (e.g., all alphas are zero) in case the data are insufficiently informative about the parameters of interest. Reduction of false discoveries is achieved first by shrinking estimates toward the prior. When there is no information in the data, the alpha point estimate is the prior mean and there are no false discoveries. As empirical evidence accumulates, posterior beliefs migrate away from the prior toward the OLS alpha estimate. In the process, discoveries begin to emerge, though they remain dampened relative to OLS. In the large-data limit, Bayesian beliefs converge on OLS with no MT correction, which is justified because in the limit there are no false discoveries. In other words, the prior embodies a particularly flexible form of conservatism—the Bayesian model decides how severe of an MT correction to make based on the informativeness of the data.

The second key model feature is the hierarchical structure that captures factors' joint behavior. Modeling factors jointly means that each alpha is shrunk toward its cluster mean (i.e., toward related factors), in addition to being shrunk toward the prior of zero. So, if we observe a cluster of factors in which most perform poorly, then this evidence reduces the posterior alpha even for the few factors with strong performance—another form of Bayesian MT correction. In addition to this Bayesian discovery control coming through shrinkage of the posterior mean alpha, the Bayesian confidence interval also plays an important role and changes as a function of the data. Indeed, having data on related factors leads to a contraction of the confidence intervals in our joint Bayesian model. So while alpha shrinkage often has the effect of reducing discoveries, the increased precision from joint estimation has the opposite effect of enhancing statistical power and thus increases discoveries.

In summary, a typical implementation of frequentist MT corrections estimates parameters independently for each factor and leaves these parameters unchanged, but inflates *p*-values to reduce the number of discoveries. In contrast, our hierarchical model leverages dependence in the data to efficiently learn about all alphas simultaneously. All data therefore help to determine the

[20] See Greenland and Robins (1991), Berry and Hochberg (1999), Efron and Tibshirani (2002), Gelman, Hill, and Yajima (2012), among others. See Gelman (2016) for an intuitive, informal discussion of the topic.

center and width of each alpha's confidence interval (Propositions 3 and 4). This leads to more precise estimates with "built-in" Bayesian MT correction.

C.2. Empirical Bayes Estimation

Given the central role of the prior, it might seem problematic that the severity of the Bayesian MT adjustment is at the discretion of the researcher. A powerful (and somewhat surprising) aspect of a hierarchical model is that the prior can be learned in part from the data. This idea is formalized in the idea of "empirical Bayes (EB)" estimation, which has emerged as a major toolkit for navigating MT in high-dimensional statistical settings (Efron (2012)).

The general approach to EB is to specify a multilevel hierarchical model and then use the dispersion in estimated effects within each level to learn about the prior parameters for that level. In our setting, the specific implementation of EB is dictated by Proposition 4. We first compute each factor's abnormal return, $\hat{\alpha}$, as the intercept in a CAPM regression on the market excess return. We then set the overall alpha prior mean, α^o, to zero to enforce conservatism in our inferences.

From here, the benefits of EB kick in. The realized dispersion in alphas across factors helps determine the appropriate prior beliefs (i.e., the appropriate values for τ_c^2, τ_s^2, and τ_w^2). For example, if we compute the average alpha for each cluster, $\hat{c}^j$ (e.g., the average value alpha, the average momentum alpha), the cross-sectional variation in $\hat{c}^j$ suggests that $\tau_c^2 \cong \frac{1}{J-1}\sum_{j=1}^{J}(\hat{c}^j - \hat{c}^{\cdot})^2$. The same idea applies to τ_s^2. Likewise, variation in observed alphas after accounting for hierarchical connections is informative about $\tau_w^2 \cong \frac{1}{NK-N-J}\sum_{i=1}^{N}(\hat{w}^i)^2$, where $\hat{w} = \hat{\alpha} - M\hat{c} - Z\hat{s}$.

The above variances illustrate that EB can help calibrate prior variances using the data itself. But those calculations are too crude, because they ignore sampling variation coming from the noise in returns, ε, which has covariance matrix Σ. EB estimates the prior variances by maximizing the prior likelihood function of the observed alphas, $\hat{\alpha} \sim N(0, \Omega(\tau_c, \tau_s, \tau_w) + \hat{\Sigma}/T)$, where the notation emphasizes that Ω depends on τ_c, τ_s, and τ_w according to (23). The likelihood function accounts for sampling variation through a plug-in estimate of the covariance matrix of factor return shocks, $\hat{\Sigma}$.[21] We collect the resulting hyperparameters in τ, that is, τ_c, τ_s, τ_w, $\hat{\Sigma}$, and β^i.

C.3. Bayesian FDR and FWER

With the EB estimates (τ) in hand, we can compute the posterior distribution of the alphas from Proposition 4. From the posterior, we can in turn compute Bayesian versions of the FDR and FWER. Suppose that we consider a factor to be "discovered" if its z-score is greater than the critical value $\bar{z} = 1.96$,

$$\frac{E(\alpha^i|\widehat{\alpha}^1,\cdots,\widehat{\alpha}^N,\tau)}{\sqrt{\mathrm{Var}(\alpha^i|\widehat{\alpha}^1,\cdots,\widehat{\alpha}^N,\tau)}} \geq \bar{z}. \tag{26}$$

[21] We provide details on our EB estimation procedure in Appendix B.

Equivalently, factor i is discovered if p-null$_i \leq 2.5\%$,[22] where we use the posterior to compute

$$\text{p-null}_i = Pr(\alpha^i \leq 0 | \hat{\alpha}^1, \cdots, \hat{\alpha}^N, \tau). \tag{27}$$

In words, p-null$_i$ is the posterior probability that the null hypothesis is true, which is the Bayesian version of a frequentist p-value. Put differently, it is the posterior probability of a "false discovery," that is, the probability that the true alpha is actually nonpositive.

We can further compute the Bayesian FDR as

$$\text{FDR}^{\text{Bayes}} = E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}} \middle| \hat{\alpha}^1, \ldots, \hat{\alpha}^N, \tau \right), \tag{28}$$

where we condition on the data including at least one discovery (so the denominator is not zero); otherwise, FDR is set to zero (see Benjamini and Hochberg (1995)).

The following proposition is a novel characterization of the Bayesian FDR, and shows that it is the posterior probability of a false discovery, averaged across all discoveries:

PROPOSITION 5 (Bayesian FDR): *Conditional on the parameters of the prior distribution τ and data with at least one discovery, the Bayesian FDR can be computed as*

$$FDR^{Bayes} = \frac{1}{\#discoveries} \sum_{i\, discovery} p\text{-}null_i \tag{29}$$

and is bounded, $\text{FDR}^{Bayes} \leq 2.5\%$.

This result shows explicitly how the Bayesian framework controls the FDR without the need for additional MT adjustments.[23] The definition of a discovery ensures that at most 2.5% of the discoveries are false according to the Bayesian posterior, which is exactly the right distribution for assessing discoveries from the perspective of the Bayesian. Further, if many of the discovered factors are highly significant (as is the case in our data), then the Bayesian FDR is much lower than 2.5%.[24]

We can also compute a Bayesian version of the FWER, which is the probability of making one or more false discoveries in total:

$$\text{FWER}^{\text{Bayes}} = Pr\left(\sum_i 1_{\{i \text{ false discovery}\}} \geq 1 \middle| \hat{\alpha}^1, \ldots, \hat{\alpha}^N, \tau \right). \tag{30}$$

[22] We use a critical value of 2.5% rather than 5% because the 1.96 cutoff corresponds to a two-sided test, while false discoveries are only on one side in the Bayesian framework.

[23] Efron (2007) includes related analysis but, to our knowledge, this particular result is new.

[24] Proposition 5 formalizes the argument of Greenland and Robins (1991) that "from the empirical-Bayes or Bayesian perspective, multiple comparisons are not really a 'problem.' Rather, the multiplicity of comparisons provides an opportunity to improve our estimates through judicious use of any prior information (in the form of model assumptions) about the ensemble of parameters being estimated."

If we define a discovery as in (26) using the standard critical value $\bar{z} = 1.96$, then we do not necessarily control the $\text{FWER}^{\text{Bayes}}$, which is a harsh criterion that is concerned with the risk of a single false discovery without regard for the number of missed discoveries. Because $\text{FWER}^{\text{Bayes}}$ is a probability that can be computed from the posterior, it is straightforward to choose a critical value $\bar{z}$ to ensure $\text{FWER}^{\text{Bayes}} \leq 5\%$ or any other level one prefers. The main point is that the Bayesian approach to replication lends itself to any inferential calculation the researcher desires because the posterior is a complete characterization of Bayesian beliefs about model parameters.

C.4. A Comparison of Frequentist and Bayesian False Discovery Control

We illustrate the benefits of Bayesian inference for our replication analysis via simulation. We assume a factor-generating process based on the hierarchical model above and, for simplicity, consider a single region (as in our empirical U.S.-only analysis), removing s^n and τ_s^2 from equations (21) and (23). We analyze discoveries as we vary the prior variances τ_c and τ_w. The remaining parameters are calibrated to our estimates for the U.S. region in our empirical analysis below.

We simulate an economy with 130 factors in 13 different clusters of 10 factors each, observed monthly over 70 years. We assume that the mean alpha, α^o, is zero. We then draw a cluster alpha from $c^j \sim N(0, \tau_c^2)$ and a factor-specific alpha as $w^i \sim N(0, \tau_w^2)$. Based on these alphas, we generate realized returns by adding Gaussian noise.[25]

We compute p-values separately using OLS with no adjustment or OLS adjusted using the Benjamini and Yekutieli (2001) (BY) method. We also use EB to estimate the posterior alpha distribution, treating τ_c and τ_w as known to simplify simulations and focus on the Bayesian updating. For OLS and BY, a discovery occurs when the alpha estimate is positive and the two-sided p-value is below 5%. For EB, we consider it a discovery when the posterior probability that alpha is negative is less than 2.5%. For each τ_c and τ_w pair, we draw 10,000 simulated samples and report average discovery rates over all simulations.

Figure 3 reports alpha discoveries based on the OLS, BY, and EB approaches. For each method, we report the true FDR in the top panels (we know the truth since this is a simulation) and the "true discovery rate"[26] in the bottom panels.

When idiosyncratic variation in true alphas is small (left panels with $\tau_w = 0.01\%$) and the variation in cluster alphas is also small (values of τ_c near zero

[25] The noise covariance matrix has a block structure calibrated to our data, with a correlation of 0.58 among factors in the same cluster and a correlation of 0.02 across clusters. The residual volatility for each factor is 10% per annum.

[26] We define the true discovery rate to be the number of significantly positive alphas according to, respectively, OLS, BY, and EB divided by the number of truly positive alphas. Given our simulation structure, half of the alphas are expected to be positive in any simulation. Some of these will be small (i.e., economically insignificant) positives, so a testing procedure would require a high degree of statistical power to detect them. This is why the true discovery rate is below one even for high values of τ_c.

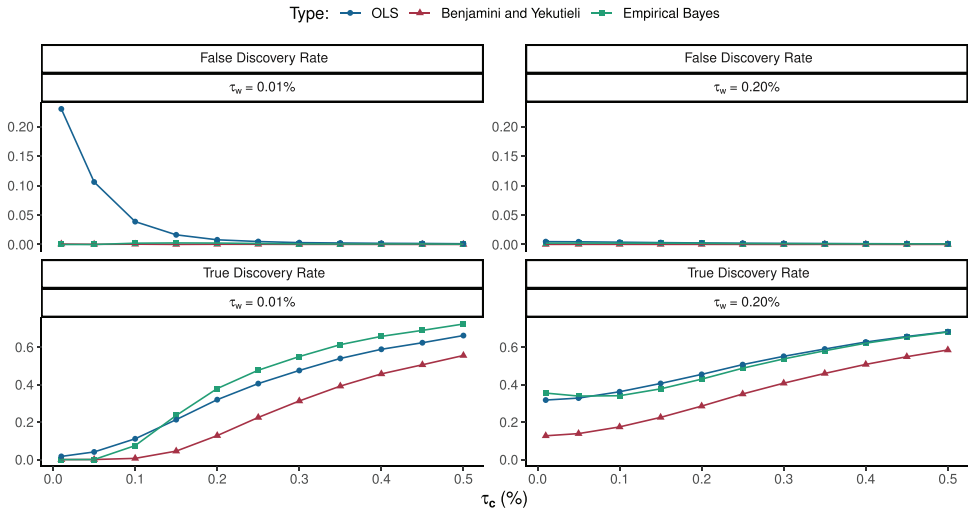

Figure 3. Simulation comparison of false discovery rates. The upper panels show the realized FDR computed as the proportion of discovered factors for which the true alpha is negative, averaged over 10,000 simulations. The lower panels show the true discovery rate computed as the number of discoveries for which the true alpha is positive divided by the total number of factors for which the true alpha is positive. The left and right panels use low and high values of idiosyncratic variation in alphas (τ_w), respectively. The x-axis varies cluster alpha dispersion, τ_c. (Color figure can be viewed at wileyonlinelibrary.com)

on the horizontal axis), alphas are very small and true discoveries are unlikely. In this case, the OLS FDR can be as high as 25% as seen in the upper left panel. However, both BY and EB successfully correct this problem and lower the FDR. The lower left panel shows that the BY correction pays a high price for its correction in terms of statistical power when τ_c is larger. In contrast, EB exhibits much better power to detect true positives while maintaining a similar false discovery control as BY. In fact, when there are more discoveries to be made in the data (as τ_c increases), EB becomes even more likely to identify true positives than OLS. This is due to the joint nature of the Bayesian model, whose estimates are especially precise compared to OLS due to EB's ability to learn more efficiently from dependent data. This result illustrates the observation by Greenland and Robins (1991) that "Unlike conventional multiple comparisons, empirical-Bayes and Bayes approaches will alter and can improve point estimates and can provide more powerful tests and more precise (narrower) interval estimators." When the idiosyncratic variation is larger ($\tau_w = 0.20\%$), there are many more true discoveries to be made, so the FDR tends to be low even for OLS with no correction. Yet, in the lower right panel, we continue to see the costly loss of statistical power suffered by the BY correction.

In summary, EB accomplishes a flexible MT adjustment by adapting to the data-generating process. When discoveries are rare so that there is a comparatively high likelihood of false discovery, EB imposes heavy shrinkage and behaves similarly to the conservative BY correction. In this case, the benefit of

conservatism costs little in terms of power exactly because true discoveries are rare. Yet, when discoveries are more likely, EB behaves more like uncorrected OLS, giving it high power to detect discoveries and suffering little in terms of false discoveries because true positives abound.

The limitations of frequentist MT corrections are well studied in the statistics literature. Berry and Hochberg (1999) note that "these procedures are very conservative (especially in large families) and have been subjected to criticism for paying too much in terms of power for achieving (conservative) control of selection effects." The reason is that, while inflating confidence intervals and *p*-values reduces the discovery of false positives, it also reduces power to detect true positives.

Much of the discussion around MT adjustments in the finance literature fails to consider the loss of power associated with frequentist corrections. But as Greenland and Hofman (2019) point out, this trade-off should be a first-order consideration for a researcher navigating MT, and frequentist MT corrections tend to place an implicit cost on false positives that can be unreasonably large. Unlike some medical contexts for example, there is no obvious motivation for asymmetric treatment of false positives and missed positives in factor research. The finance researcher may be willing to accept the risk of a few false discoveries to avoid missing too many true discoveries. In statistics, this is sometimes discussed in terms of an (abstract) cost of Type I versus Type II errors,[27] but in finance we can make this cost concrete: We can look at the profit of trading on the discovered factors, where the cost of false discoveries is then the resulting extra risk and money lost (Section III.C.1).

II. A New Public Data Set of Global Factors

We study a global data set with 153 factors in 93 countries. In this section, we provide a brief overview of our data construction. We have posted the data and code along with extensive documentation detailing each implementation choice that we make for each factor.[28]

A. *Factors*

The set of factors that we study is based on the exhaustive list compiled by Hou, Xue, and Zhang (2020). They study 202 different characteristic signals from which they build 452 factor portfolios. The proliferation is due to treating one-, six-, and 12-month holding periods for a given characteristic as different factors, and due to their inclusion of both annual and quarterly updates of

[27] As Greenland and Robins (1991) point out, "Decision analysis requires, in addition to the likelihood function, a loss function, which indicates the cost of each action under the various possible values for the unknown parameter (benefits would be expressed as negative costs). Construction of a loss function requires one to quantify costs in terms of dollars, lives lost, or some other common scale."

[28] The data and code are available at https://jkpfactors.com/ and https://github.com/bkelly-lab/ReplicationCrisis. The data will be updated over time and will also be available via WRDS.

some accounting-based factors. In contrast, we focus on a one-month holding period for all factors, and we only include the version that updates with the most recent accounting data—this could be either annual or quarterly. Finally, we exclude a small number of factors for which data are not available globally. This gives us a set of 180 feasible global factors. For this set, we exclude factors based on industry or analyst data because they have comparatively short samples.[29] This leaves us with 138 factors. Finally, we add 15 factors studied in the literature that were not included in Hou, Xue, and Zhang (2020). For each characteristic, we build the one-month-holding-period factor return within each country as follows. First, in each country and month, we sort stocks into characteristic terciles (top/middle/bottom third) with breakpoints based on non-micro stocks in that country.[30] For each tercile, we compute its "capped value weight" return, meaning that we weight stocks by their market equity winsorized at the NYSE 80$^{\text{th}}$ percentile. This construction ensures that tiny stocks have tiny weights and any one mega stock does not dominate a portfolio in an effort to create tradable, yet balanced, portfolios.[31] The factor is then defined as the high-tercile return minus the low-tercile return, corresponding to the excess return of a long-short zero-net-investment strategy. The factor is long (short) the tercile identified by the original paper to have the highest (lowest) expected return.

We scale all factors such that their monthly idiosyncratic volatility is $10\%/\sqrt{12}$ (i.e., 10% annualized), which ensures cross-sectional stationarity and a prior that factors are similar in terms of their information ratio. Finally, we compute each factor's $\hat{\alpha}^i$ via an OLS regression on a constant and the corresponding region's market portfolio.

For a factor return to be nonmissing, we require that it have at least five stocks in each of the long and short legs. We also require a minimum of 60 nonmissing monthly observations for each country-specific factor for inclusion in our sample. When grouping countries into regions (U.S., developed ex-U.S., and emerging), we use the Morgan Stanley Capital International (MSCI) development classification as of January 7, 2021. When aggregating factors across countries, we use capitalization-weighted averages of the country-specific factors. For the developed and emerging market factors, we require that at least three countries have nonmissing factor returns.

[29] Global industry codes (GICS) are only available from 2000. I/B/E/S data are available from the mid-1980s but coverage in early years is somewhat sparse.

[30] Specifically, we start with all nonmicro stocks in a country (i.e., larger than NYSE 20$^{\text{th}}$ percentile) and sort them into three groups of equal numbers of stocks based on the characteristic, say book-to-market. We then distribute the micro-cap stocks into the three groups based on the same characteristic breakpoints. This process ensures that the nonmicro stocks are distributed equally across portfolios, creating more tradable portfolios.

[31] For robustness, Figure IA.1 of the Internet Appendix reports our replication results when using standard, uncapped value weights to construct factors.

B. Clusters

We group factors into clusters using hierarchical agglomerative clustering (Murtagh and Legendre (2014)). We define the distance between factors as one minus their pairwise correlation and use the linkage criterion of Ward (1963). The correlation is computed based on CAPM-residual returns of U.S. factors signed as in the original paper. Figure IA.15 of the Internet Appendix shows the resulting dendrogram, which illustrates the hierarchical clusters identified by the algorithm. Based on the dendrogram, we choose 13 clusters that demonstrate a high degree of economic and statistical similarity. The cluster names indicate the types of characteristics that dominate each group: Accruals*, Debt Issuance*, Investment*, Leverage*, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size*, Short-Term Reversal, and Value, where (*) indicates that these factors bet against the corresponding characteristic (e.g., accrual factors go long stocks with low accruals while shorting those with high accruals). Figure IA.16 shows that the average within-cluster pairwise correlation is above 0.5 for nine out of 13 clusters. Table IA.II provides details on the cluster assignment, sign convention, and original publication source for each factor.

C. Data and Characteristics

Return data are from CRSP for the United States (beginning in 1926) and from Compustat for all other countries (beginning in 1986 for most developed countries).[32] All accounting data are from Compustat. For international data, all variables are measured in U.S. dollars (based on exchange rates from Compustat) and excess returns are relative to the U.S. Treasury bill rate. To alleviate the influence of data errors in the international data, we winsorize returns from Compustat at 0.1% and 99.9% each month.

We restrict our focus to common stocks that are identified by Compustat as the primary security of the underlying firm and assign stocks to countries based on the country of their exchange.[33] In the United States, we include delisting returns from CRSP. If a delisting return is missing and the delisting is for a performance-based reason, we set the delisting return to −30% following Shumway (1997). In the global data, delisting returns are not available, so all performance-based delistings are assigned a return of −30%.

We build characteristics in a consistent way, that sometimes deviates from the exact implementation used in the original reference. For example, for characteristics that use book equity, we always follow the method in Fama and French (1993). Furthermore, we always use the most recent accounting data, whether annual or quarterly. Quarterly income and cash flow items are aggregated over the previous four quarters to avoid distortions from seasonal

[32] Table IA.IV shows start date and other information for all countries included in our data set.

[33] Compustat identifies primary securities in the United States, Canada, and rest of the world. This means that some firms can have up to three securities in our data set. In practice, the vast majority of firms (97%) only have one security in our sample at a given point in time.

effects. We assume that accounting data are available four months after the fiscal period end. When creating valuation ratios, we always use the most recent price data following Asness and Frazzini (2013). Section IX in the Internet Appendix contains detailed documentation of our data set.

D. EB Estimation

We estimate the hyperparameters and the posterior alpha distributions of our Bayesian model via EB. Appendix B provides details on the EB methodology and the estimated parameters.

III. Empirical Assessment of Factor Replicability

We now report replication results for our global factor sample. We first present an internal validity analysis by studying U.S. factors over the full sample. We then analyze external validity in the global cross section and in the time series (postpublication factor returns).

A. Internal Validity

We plot the full-sample performance of U.S. factors in Figure 4. Each panel shows the CAPM alpha point estimate of each factor corresponding to the dot at the center of the vertical bars. Vertical bars represent the 95% confidence interval for each estimate. Bar colors and linetypes differentiate between three types of factors. Solid blue indicates factors that are significant in the original study and remain significant in our full sample. Dashed red indicates factors that are significant in the original study but insignificant in our test. Dotted green indicates factors that are not significant in the original study but are included in the sample of Hou, Xue, and Zhang (2020).

The four panels in Figure 4 differ in how the alphas and their confidence intervals are estimated. The upper left panel reports the simple OLS estimate of each alpha, $\hat{\alpha}_{\text{ols}}$, and the 95% confidence intervals based on unadjusted standard errors, $\hat{\alpha}_{\text{ols}} \pm 1.96 \times SE_{\text{ols}}$.[34] The factors are sorted by OLS $\hat{\alpha}$ estimate, and we use this ordering for the other three panels as well. We find that the OLS replication rate is 82.4%, computed as the number of solid blue factors (98) divided by the sum of solid blue and dashed red factors (119). Based on OLS tests, factors are highly replicable.

The upper right panel repeats this analysis using the MT adjustment of BY, which is advocated by Harvey, Liu, and Zhu (2016) and implemented by Hou, Xue, and Zhang (2020). This method leaves the OLS point estimate unchanged, but inflates the p-value. We illustrate this visually by widening the alpha confidence interval. Specifically, we find the BY-implied critical value[35] in our

[34] We define SE_{ols} as the diagonal of the alpha covariance matrix $\hat{\Sigma}$, which we estimate according to Appendix B.

[35] We compute the BY-implied critical value as the average of the t-statistic of the factor that is just significant based on BY (the factor with the highest BY-adjusted p-value below 5%) and the

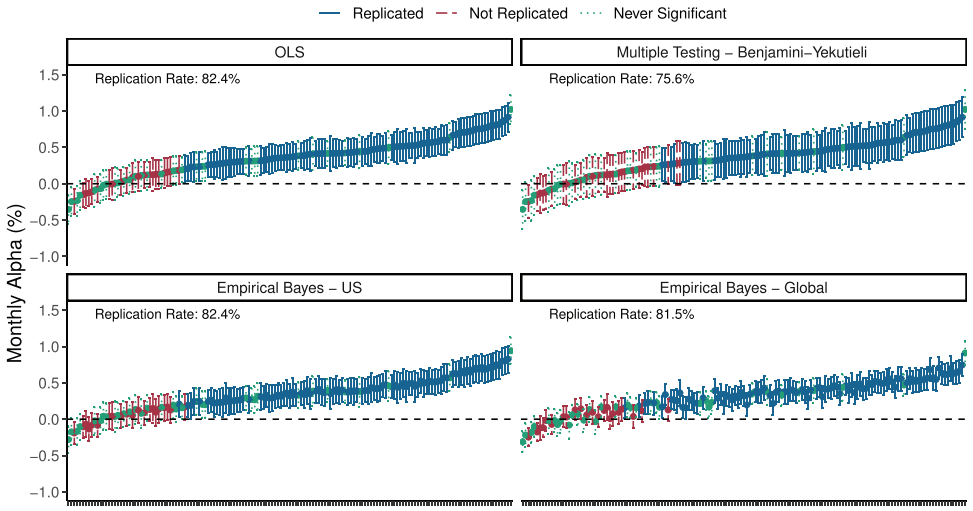

Figure 4. Alpha distributions for U.S. factors. The figure shows point estimates and confidence intervals for U.S. factors. The upper left panel depicts OLS estimates. The upper right panel uses the OLS point estimate but adjusts the confidence interval following the BY procedure. The lower left panel shows our EB posterior confidence intervals using only U.S. data. The lower right panel continues to show EB results for U.S. factors, but estimates the U.S. factor posterior from global data rather than U.S.-only data. Solid blue (dashed red) confidence intervals correspond to factors that were significant in the original study and that we find to be significant (insignificant) based on the method in each panel. Dotted green intervals correspond to factors that the original study find to be insignificant or that the original study does not evaluate in terms of average return significance. The order of factors is the same in all panels and is arranged from lowest OLS alpha to highest. Table IA.III lists the factor names arranged in the same order. (Color figure can be viewed at wileyonlinelibrary.com)

sample to have a t-statistic of 2.7, and we compute the corresponding confidence interval as $\hat{\alpha}_{\text{ols}} \pm 2.7 \times SE_{\text{ols}}$. We deem a factor as significant according to the BY method if this interval lies entirely above zero. Naturally, this widening of confidence intervals produces a lower replication rate of 75.6%. However, the BY correction does not materially change the OLS-based conclusion that factors appear to be highly replicable.

The lower left panel is based on our EB estimates using the full sample of U.S. factors. For each factor, we use Proposition 4 to compute its posterior mean, $E(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}})$, shown as the dot at the center of the confidence interval. These dots change relative to the OLS estimates, in contrast to BY and other frequentist MT methods that only change the size of the confidence intervals. We also compute the posterior volatility to produce Bayesian confidence intervals, $E(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}}) \pm 1.96 \times \sigma(\alpha_i|(\hat{\alpha}_j)_{j \text{ any US factor}})$. The replication rate based on Bayesian model estimates is 82.4%, larger than BY and, coincidentally, the same as the OLS replication rate. This replication rate has

t-statistic of the factor that is just insignificant (the factor with the lowest BY-adjusted p-value above 5%).

a built-in conservatism from the zero-alpha prior, and it further accounts for the multiplicity of factors because each factor's posterior depends on *all* of the observed evidence in the United States (not just own-factor performance).

The lower right panel again reports EB estimates for U.S. factors, but now we allow the posterior to depend on data from all over the world, not just on U.S. data. That is, we compute the posterior mean and variance for each U.S. factor conditional on the alpha estimates for all factors in all regions. The resulting replication rate is 81.5%, which is slightly lower than the EB replication rate using only U.S. data. Some posterior means are reduced due to the fact that some factors have not performed as well outside the United States, which affects posterior means for the United States through the dependence among global alphas. For example, when the Bayesian model seeks to learn the true alpha of the "U.S. change in book equity" factor, the Bayesian's conviction regarding positive alpha is reduced by accounting for the fact that the international version of this factor has underperformed the U.S. version.[36]

To further assess internal validity, we investigate the replication rate for U.S. factors when those factors are constructed from subsamples based on stock size. One of the leading criticisms of factor research replicability is that results are driven by illiquid small stocks whose behavior largely reflects market frictions and microstructure as opposed to just economic fundamentals or investor preferences. In particular, Hou, Xue, and Zhang (2020) argue that they find a low replication rate because they limit the influence of micro-caps. We find that factors demonstrate a high replication rate throughout the size distribution. Panel A of Figure 5 summarizes replication rates for U.S. size categories shown in the five bars: mega stocks (largest 20% of stocks based on NYSE breakpoints), large stocks (market capitalization between the 80th and 50th percentile of NYSE stocks), small stocks (between the 50th and 20th percentile), micro stocks (between the 20th and 1st percentile), and nano stocks (market capitalization below the 1st percentile).

We see that the EB replication rates in mega- and large-stock samples are 77.3% and 79.8%, respectively. This is only marginally lower than the overall U.S. sample replication rate of 82.4%, indicating that criticisms of factor replicability based on arguments around stock size or liquidity are largely groundless. For comparison, small, micro, and nano stocks deliver replication rates of 85.7%, 85.7%, and 68.1%, respectively.

In Panel B of Figure 5, we provide U.S. factor replication rates by theme cluster. Eleven out of 13 themes are replicable with a rate of 50% or better, with the exceptions being the low leverage and size themes. To understand these exceptions, we note that size factors are stronger in emerging markets (bottom panel of Figure IA.7) and among micro and nano stocks (bottom panels of Figure IA.8). The theoretical foundation of the size effect is compensation

[36] To provide a few more details on this example, the U.S. factor based on annual change in book equity (be_gr1a) has a posterior volatility of 0.095% using U.S. data and 0.077% using global data, leading to a tighter confidence interval with the global data. However, the posterior mean is 0.22% using only U.S. data and 0.13% using global data.

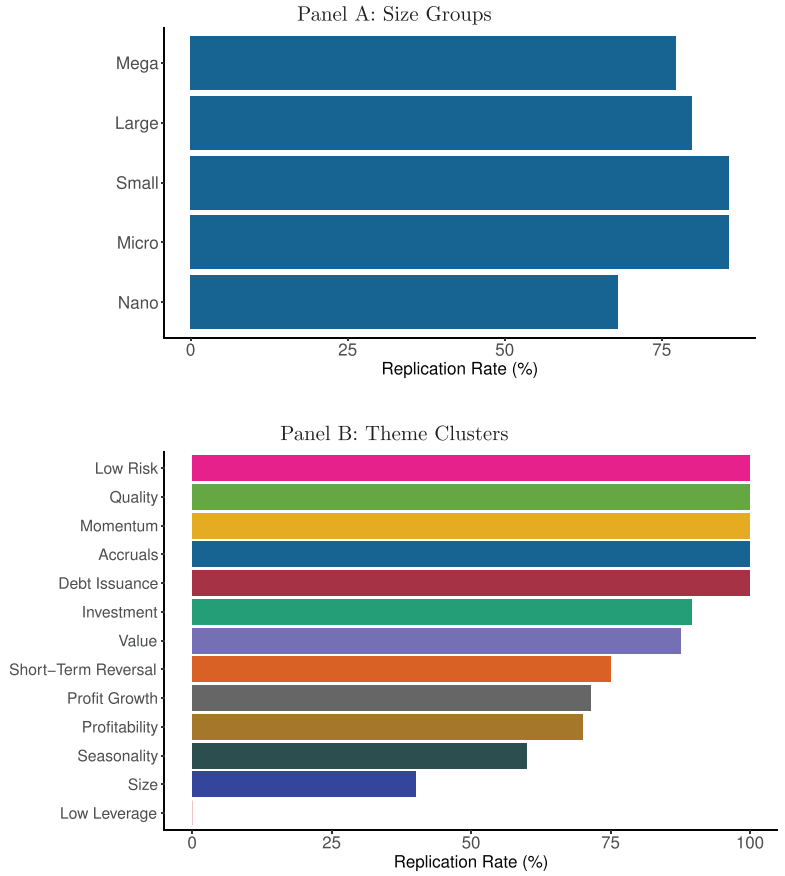

Figure 5. U.S. replication rates by size group and theme cluster. Panel A summarizes replication rates for U.S. factors formed from subsamples defined by stocks' market capitalization using our EB method. Panel B shows replication rates for U.S. factors in each theme cluster. (Color figure can be viewed at wileyonlinelibrary.com)

for market illiquidity (Amihud and Mendelson (1986)) and market liquidity risk (Acharya and Pedersen (2005)). Theory predicts that the illiquidity (risk) premium should be the same order of magnitude as the differences in trading costs, and these differences are simply much larger in emerging markets and among micro stocks.

Another reason some factors and themes appear insignificant is that we do not account for other factors. Factors published after 1993 are routinely benchmarked to the Fama-French three-factor model (and, more recently, to the updated five-factor model). Some factors are insignificant in terms of raw return or CAPM alpha, but their alpha becomes significant after controlling for other factors. Indeed, this explanation accounts for the lack of replicability for the low-leverage theme. While CAPM alphas of low-leverage factors are

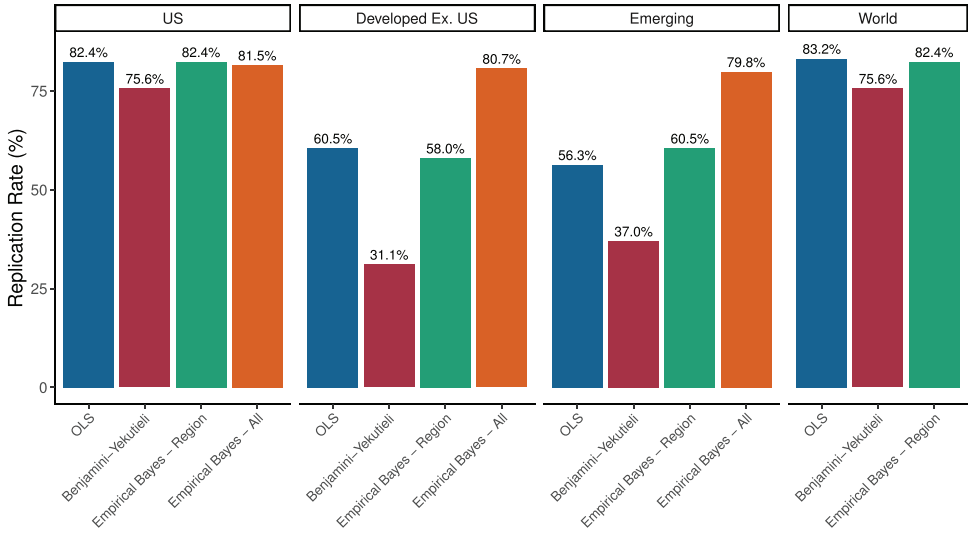

Figure 6. Replication rates in global data. We summarize replication rates for factors in three regions (U.S., developed ex.-U.S., and emerging) and for the world as a whole. A factor in a given region is the capitalization-weighted average factor for countries in that region. We report OLS replication rates with no adjustment and with MT adjustment of BY. We also report replication rates based on the EB posterior. We consider two EB methods. In both methods, the replication rate corresponds only to factors within the region of interest, but the posterior is computed by conditioning either on data from that region alone ("Empirical Bayes – Region") or on the full global sample ("Empirical Bayes – All"). We deem a factor successfully replicated if its 95% confidence interval excludes zero for a given method. (Color figure can be viewed at wileyonlinelibrary.com)

insignificant, we find that it is one of the best-performing themes when we account for multiple factors (see Section III.D.2 below).

B. External Validity

We find a high replication rate in our full-sample analysis, indicating that the large majority of factors are reproducible at least in-sample. We next study the external validity of these results in international data and in postpublication U.S. data.

B.1. Global Replication

Figure 6 shows corresponding replication rates around the world. We report replication rates from four testing approaches: (i) OLS with no adjustment, (ii) OLS with MT adjustment of BY, (iii) the EB posterior conditioning only on factors within a region ("Empirical Bayes – Region"), and (iv) EB conditioning on factors in all regions ("Empirical Bayes – All"). Even when using all global data to update the posterior of all factors, the reported Bayesian replication rate applies only to the factors within the stated region.

The first set of bars establishes a baseline by showing replication rates for the U.S. sample, summarizing the results from Figure 4. The next two sets of bars correspond to the developed ex.-U.S. sample and the emerging markets sample, respectively.[37] Each region's factor is a capitalization-weighted average of that factor among countries within a given region, and the replication rate describes the fraction of significant CAPM alphas for these regional factors.

OLS replication rates in developed and emerging markets are generally lower than in the United States, and the frequentist BY correction has an especially large negative impact on replication rate. This is a case in which the Bayesian approach to MT is especially powerful. Even though the alphas of all regions are shrunk toward zero, the global information set helps EB achieve a high degree of precision, narrowing the posterior distribution around the shrunk point estimate. We can see this in increments. First, the EB replication rate using region-specific data ("Empirical Bayes – Region" in the figure) is just below the OLS replication rate but much higher than the BY rate. When the posterior leverages global data ("Empirical Bayes – All" in the figure), the replication rate is even higher, reflecting the benefits of sharing information across regions, as recommended by the dependence among alphas in the hierarchical model.

Finally, we use the global model to compute, for each factor, the capitalization-weighted average alpha across all countries in our sample ("World" in the figure). Using data from around the world, we find a Bayesian replication rate of 82.4%.

In summary, our EB-All method yields a high replication rate in all regions. That said, the OLS replication rates are lower outside the U.S. than in the U.S., which is primarily due to the fact that foreign markets have shorter time samples—the point estimates of alphas are similar in magnitude for the U.S. and international data. Figure 7 shows the alpha of each U.S. factor against the alpha of the corresponding factor for the world ex.-U.S. universe. The data cloud aligns closely with the 45° line, demonstrating the close similarity of alpha magnitudes in the two samples. But shorter international samples widen confidence intervals, and this is the primary reason for the drop in OLS replication rates outside the United States.

B.2. Time-Series Out-of-Sample Evidence

McLean and Pontiff (2016) document the intriguing fact that, following publication, factor performance tends to decay. They estimate an average postpublication decline of 58% in factor returns. In our data, the average in-sample alpha is 0.49% per month and the average out-of-sample alpha is 0.26% when looking postoriginal sample, implying a decline of 47%.

[37] The developed and emerging samples are defined by the MSCI development classification and include 23 and 27 countries, respectively. The remaining 43 countries in our sample that are classified as neither developed nor emerging by MSCI do not appear in our developed and emerging region portfolios, but they are included in the "world" versions of our factor portfolios.

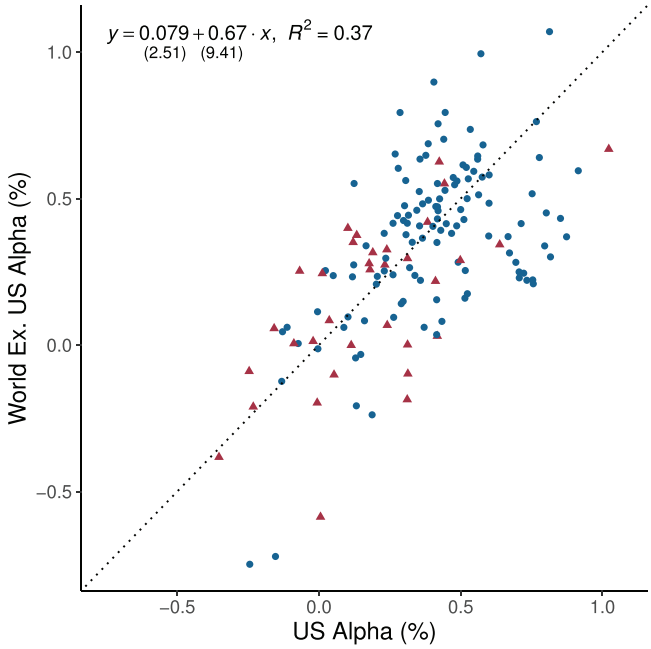

Figure 7. U.S. factor alphas versus world ex.-U.S. The figure compares OLS alphas for U.S. factors versus their international counterpart. Each world ex.-U.S. factor is a capitalization-weighted average of the factor in all other countries of our sample. Blue circles correspond to factors that were significant in the original study, while red triangles are those for which the original paper did not find a significant effect (or did not study the factor in terms of average return significance). The dotted line is the 45º line. The figure also shows a regression of world ex.-U.S. alpha on U.S. alpha. (Color figure can be viewed at wileyonlinelibrary.com)

We gain further economic insight by looking at these findings cross-sectionally. Figure 8 provides a cross-sectional comparison of the in-sample and out-of-sample alphas of our U.S. factors. The in-sample period is the sample studied in the original reference. The out-of-sample period in Panel A is the period before the start of in-sample period, while in Panel B it is the period following the in-sample period. Panel C defines out-of-sample as the combined data from the periods before and after the originally studied sample. We find that 82.6% of the U.S. factors that were significant in the original publication also have positive returns in the preoriginal sample, 83.3% are positive in the postoriginal sample, and 87.4% are positive in the combined out-of-sample period. When we regress out-of-sample alphas on in-sample alphas using generalized least squares (GLS), we find a slope coefficient of 0.57, 0.26, and 0.35 in Panels A, B, and C, respectively. The slopes are highly significant (ranging from $t = 3.5$ to $t = 5.3$), indicating that in-sample alphas contain something "real" rather than being the outcome of pure data mining, as factors that performed better in-sample also tend to perform better out-of-sample.

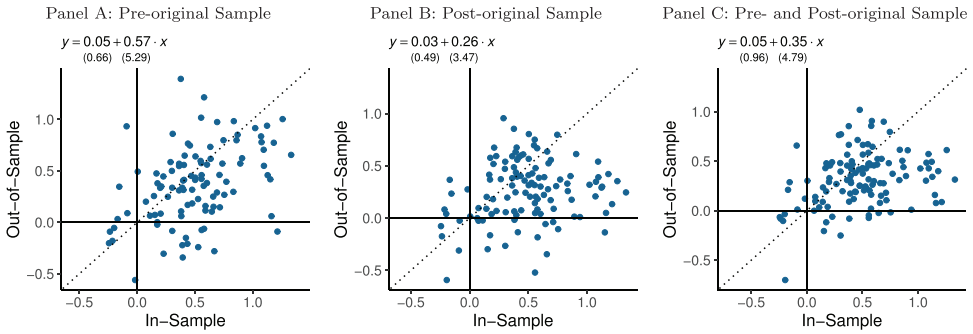

Figure 8. In-sample versus out-of-sample alphas for U.S. factors. The figure shows OLS alphas for U.S. factors during the in-sample period (i.e., the period studied in the original publication) versus out-of-sample alphas. In Panel A, out-of-sample is the period before the in-sample period. In Panel B, out-of-sample is the period after the in-sample period. In Panel C, out-of-sample includes both the period before and the period after the in-sample period. We require at least five years of out-of-sample data for a factor to be included, amounting to 102, 115, and 119 factors in Panel A, B, and C. The figure also shows feasible GLS estimates of out-of-sample alphas on in-sample alphas. To implement feasible GLS, we assume that the error variance-covariance matrix is proportional to the full-sample CAPM residual variance-covariance matrix, $\hat{\Sigma}/T$, described in Appendix B. The dotted line is the 45º line. (Color figure can be viewed at wileyonlinelibrary.com)

The significantly positive slope allows us to reject the hypothesis of "pure alpha-hacking," which would imply a slope of zero, as seen in Proposition 1. Further, the regression intercept is positive, while alpha-hacking of the form studied in Proposition 1 would imply a negative intercept. That the slope coefficient is positive and less than one is consistent with the basic Bayesian logic of equation (4). As we emphasize in Section I, a Bayesian would expect at least some attenuation in out-of-sample performance. This is because the published studies report the OLS estimate, while Bayesian beliefs shrink the OLS estimate toward the zero-alpha prior. More specifically, with no alpha-hacking or arbitrage, the Bayesian expects a slope of approximately 0.9 using equation (5) and our EB hyperparameters (see Table I).[38] Hence, the slope coefficients in Figure 8 are too low relative to this Bayesian benchmark. In addition to the moderate slope, there is evidence that the dots in Figure 8 have a concave shape (as seen more clearly in Figure IA.3). These results indicate that, while we can rule out pure alpha-hacking (or *p*-hacking), there is some evidence that the highest in-sample alphas may be data-mined or arbitraged down.

From the Bayesian perspective, another interesting evaluation of time-series external validity is to ask whether the new information contained in out-of-sample data moves the posterior alpha toward zero. Imagine a Bayesian observing the arrival of factor data in real time. As new data arrive, she updates her beliefs for all factors based on the information in the full cross section of factor data. In the top panel of Figure 9, we show how the Bayesian's

[38] The slope is $\kappa = 1/(1 + \sigma^2/(T\tau^2)) = 0.9$, where $\sigma^2 = 10\%^2/12$, the average in-sample period length is $T = 420$ months, and $\tau^2 = \tau_c^2 + \tau_w^2 = (0.35\%)^2 + (0.21\%)^2 = (0.41\%)^2$.

Table I

Hyperparameters of the Prior Distribution Estimated by Maximum Likelihood

This table presents τ_c as the estimated dispersion in cluster alphas (e.g., the dispersion in the alpha of the value cluster alpha, momentum cluster). When we estimate a single region, τ_w is the idiosyncratic dispersion of alphas within each cluster. When we jointly estimate several regions, then τ_s is the estimated dispersion in alphas across signals within each cluster, and τ_w is the estimated idiosyncratic dispersion in alphas for factors identified by their signal and region.

Sample	τ_c	τ_w	τ_s
USA	0.35%	0.21%	
Developed	0.24%	0.18%	
Emerging	0.32%	0.24%	
USA, Developed & Emerging	0.30%	0.19%	0.10%
World	0.37%	0.23%	
World ex.-US	0.29%	0.20%	
USA—Mega	0.26%	0.16%	
USA—Large	0.31%	0.18%	
USA—Small	0.44%	0.26%	
USA—Micro	0.48%	0.32%	
USA—Nano	0.42%	0.28%	

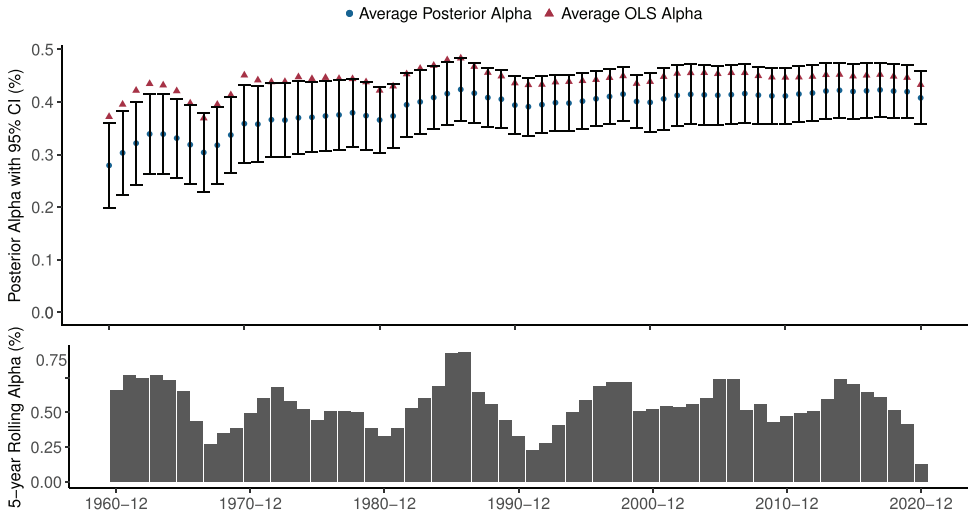

Figure 9. World factor alpha posterior distribution over time. The top panel depicts the CAPM alpha and 95% posterior confidence interval for an equal-weighted portfolio of world factors based on EB posteriors reestimated in December each year. That is, each blue circle is $E(\frac{1}{N}\sum_i \alpha^i |$ data until time $t)$ and the vertical lines are ± 2 times the posterior volatility. Red triangles show average OLS alpha at each point in time, $\frac{1}{N}\sum_i \hat{\alpha}^i_{t_0^i,t}$, estimated using data through date t. The bottom panel reports the average monthly alpha for all factors in a rolling five-year window. The results are based on factors found to be significant in the original paper with data available since 1955. (Color figure can be viewed at wileyonlinelibrary.com)

posterior of the average alpha would have evolved in real time.[39] We focus on all the world factors that are available since at least 1955 and significant in the original paper. Starting in 1960, we reestimate the hierarchical model using the EB estimator in December of each year. The plot shows the CAPM alpha and corresponding 95% confidence interval of an equal-weighted portfolio of the available factors. The posterior mean alpha becomes relatively stable from the mid-1980s, around 0.4% per month. Further, as empirical evidence has accumulated over time, the confidence interval narrows by one-third, from about 0.16% in 1960 to 0.10% in 2020.

To understand the posterior alpha, Figure 9 also shows the average OLS alpha (red triangles) and the bottom panel in Figure 9 shows the rolling five-year average monthly alpha among all these factors. We see that the EB posterior is below the OLS estimate, especially in the beginning, which occurs because the Bayesian posterior is shrunk toward the zero prior. Naturally, periods of good performance increase the posterior mean as well as the OLS estimate, and vice versa for poor performance. Over time, the OLS estimate moves nearer to the Bayesian posterior mean.

To further understand why the posterior alpha is relatively stable with a tightening confidence interval, consider the following simple example. Suppose a researcher has $T = 10$ years of data for factors with an OLS alpha estimate of $\hat{\alpha} = 10\%$ with standard error $\sigma/\sqrt{T}$. Further, assume that their zero-alpha prior is equally as informative as their 10-year sample (i.e., $\tau = \sigma/\sqrt{T}$). Then, the shrinkage factor is $\kappa = 1/2$ using equation (5). So, after observing the first 10 years with $\hat{\alpha} = 10\%$, the Bayesian expects a future alpha of $E(\alpha|\hat{\alpha}) = 5\%$ (equation (4)). What happens if this Bayesian belief is confirmed by additional data, that is, the factor realizes an alpha of 5% over the next 10 years? In this case, the full-sample OLS alpha is $\hat{\alpha} = 7.5\%$, but now the shrinkage factor becomes $\kappa = 2/3$ because the sample length doubles, $T = 20$. This results in a posterior alpha of $E(\alpha|\hat{\alpha}) = 7.5\% \cdot 2/3 = 5\%$. Naturally, when beliefs are confirmed by additional data, the posterior mean does not change. Nevertheless, we learn something from the additional data, because our conviction increases as the posterior variance is reduced. If $\sigma = 0.1$, the posterior volatility $\sqrt{\text{Var}(\alpha|\hat{\alpha})} = \sigma\sqrt{\frac{\kappa}{T}}$ goes from 2.2% with 10 years of data to 1.8% with 20 years of data, and the confidence interval, $[E(\alpha|\hat{\alpha}) \pm 2\sqrt{\text{Var}(\alpha|\hat{\alpha})}]$, decreases from $[0.5\%, 9.5\%]$ to $[1.3\%, 8.7\%]$.

C. Bayesian MT

A great advantage of Bayesian methods for tackling challenges in MT is that, from the posterior distribution, we can make explicit probability calculations

[39] Here, we keep τ_c and τ_w fixed at their full-sample values of 0.37% and 0.23% to mimic the idea of a decision maker who starts with a given prior and updates this view based on new data, while keeping the prior fixed. Figure IA.4 shows that the figure is almost the same with rolling estimates of τ_c and τ_w, and Figure IA.5 shows that this consistency arises because the rolling estimates are relatively stable.

for essentially any inferential question. We simulate from our EB posterior to investigate the FDRs and FWERs among the set of global factors that were significant in the original study. We define a false discovery as a factor where we claim that the alpha is positive, but the true alpha is negative.[40]

First, based on Proposition 5, we calculate the Bayesian FDR in our sample as the average posterior probability of a false discovery, p-null, among all discoveries. We find that $\text{FDR}^{\text{Bayes}} = 0.1\%$, meaning that we expect roughly one discovery in 1,000 to be a false positive given our Bayesian hierarchical model estimates. The posterior standard error for $\text{FDR}^{\text{Bayes}}$ is 0.3% with a confidence interval of [0,1%]. In other words, the model generates a highly conservative MT adjustment in the sense that once a factor is found to be significant, we can be relatively confident that the effect is genuine.

We can also use the posterior to make other inference calculations. We compute the FWER, which we define as the probability of at least one false discovery. We simulate 1,000,000 draws of the vector of alphas that were deemed to be discoveries from the EB posterior and compute

$$\text{FWER}^{\text{Bayes}} = \frac{1}{1{,}000{,}000} \sum_{s=1}^{1{,}000{,}000} 1_{\{n_s \geq 1\}} = 5.5\%,$$

where n_s is the number of false discoveries in simulation s. In other words, the probability of at least one alpha having the wrong sign is 5.5%. The $\text{FWER}^{\text{Bayes}}$ is naturally much higher than the $\text{FDR}^{\text{Bayes}}$ given the extreme conservatism built into the FWER's definition. Whether it is too high is subjective. A nice aspect of our approach is that a researcher can control the $\text{FWER}^{\text{Bayes}}$ as desired. For example, using a t-statistic threshold of 2.78 rather than 1.96 leads to $\text{FWER}^{\text{Bayes}} = 0.8\%$.

From the posterior, we can also compute the expected fraction of discovered factors that are "true," which in general is different than the replication rate. The replication rate is the fraction of factors having $E(\alpha_i|\text{data})/\sigma(\alpha_i|\text{data}) > 1.96$, while the expected fraction of true factors is $\frac{1}{n}\sum_i E(1_{\alpha_i>0}|\text{data}) = \frac{1}{n}\sum_i Pr(\alpha_i > 0|\text{data})$. The replication rate gives a conservative take on the number of true factors—the expected fraction of true factors is typically higher than the replication rate. To understand this conservatism, consider an example in which all factors have a 90% posterior probability of being true. These would all individually be counted as "not replicated," but they would contribute to a high expected fraction of true factors. Indeed, we estimate that the expected fraction of factors with truly positive alphas is 94% (with a posterior standard error of 1.3%), which is notably higher than our estimated replication rate.

[40] In particular, we define a discovery as a factor for which the posterior probability of the true alpha being negative is less than 2.5%. With this definition, we start with 153 world factors. We then focus on the 119 factors that were significant in the original studies. Of these 119 factors, 98 are considered discoveries.

C.1. Economic Benefits of More Powerful Tests

MT adjustments should ultimately be evaluated by whether they lead to better decisions. It is important to balance the relative costs of false positives versus false negatives, and the appropriate trade-off depends on the context of the problem (Greenland and Hofman (2019)). We apply this general principle in our context by directly measuring costs in terms of investment performance. Specifically, we can compute the difference in out-of-sample investment performance from investing using factors chosen with different methods. We compare two alternatives. One is the BY decision rule advocated by Harvey, Liu, and Zhu (2016), which is a frequentist MT method that successfully controls false discoveries relative to OLS, but in doing so sacrifices power (the ability to detect true positives). The second alternative is our EB method, whose false discovery control typically lies somewhere between BY and unadjusted OLS. EB uses the data sample itself to decide whether its discoveries should behave more similarly to BY or to unadjusted OLS.

For investors, the optimal decision rule is the one that leads to the best performance out-of-sample. For the most part, the set of discovered factors for BY and EB coincide. It is only in marginal cases where they disagree, which occurs in our sample when EB makes a discovery that BY deems insignificant. Therefore, to evaluate MT approaches in economic terms, we track the out-of-sample performance of factors included by EB but excluded by BY. If the performance of these is negative on average, then the BY correction is warranted and preferred by the investor.

We find that the out-of-sample performance of factors discovered by EB but not BY is positive on average and highly significant. The alpha for these marginal cases is 0.35% per month among U.S. factors ($t = 5.1$).[41] This estimate suggests that the BY decision rule is too conservative: An investor using the rule would fail to invest in factors that subsequently have a high out-of-sample return. Another way to see that the BY decision rule is too conservative comes from the connection between the Sharpe ratio and t-statistics: $t = \text{SR}\sqrt{T}$. If we have a factor with an annual Sharpe ratio of 0.5, an investor using the 1.96 cutoff would in expectation invest in the factor after 15 years, whereas an investor using the 2.78 cutoff would not start investing until observing the factor for 31 years.

C.2. Addressing Unobserved Factors, Publication Bias, and Other Biases

A potential concern with our replication rate is that the set of factors that make it into the literature is a selected sample. In particular, researchers may have tried many different factors, some of which are observed in the literature, while others are unobserved because they never got published. Unobserved

[41] For the developed ex.-U.S. sample, the monthly alpha for marginal cases is 0.24% per month ($t = 5.3$), and for the emerging sample it is 0.27% ($t = 3.7$), in favor of the EB decision rule. Table IA.I reports additional details for this analysis.

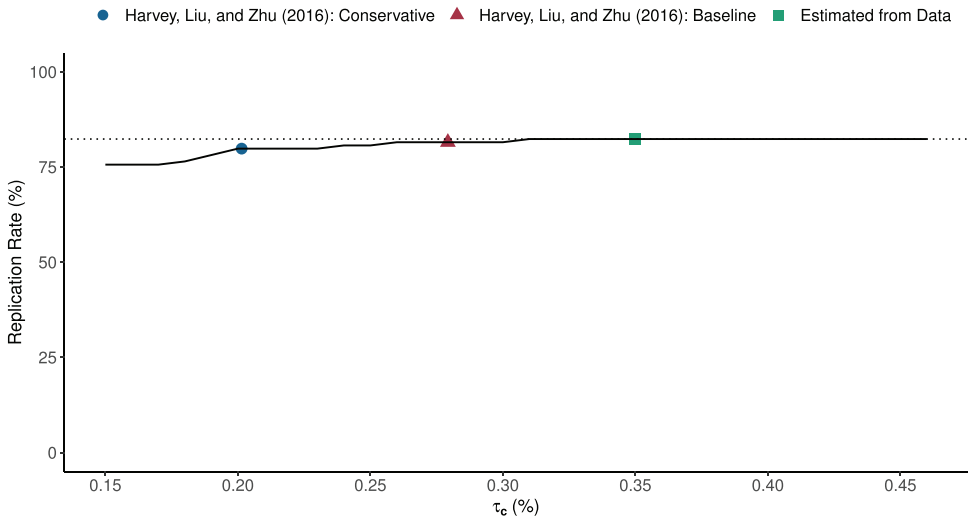

Figure 10. Replication rate with prior estimated in light of unobserved factors. The figure shows how the replication rate in the United States varies when changing the τ_c parameter. The τ_w parameter is fixed at the estimate value of 0.21%. The dotted line shows our replication rate of 82.4%. The green square highlights the value estimated in the data $\tau_c = 0.35\%$. The red triangle and the blue circle highlight values that are found by estimating the EB model according to assumptions about unobserved factors from Harvey, Liu, and Zhu (2016). The values are $\tau_c = 0.28\%$ in the baseline scenario and $\tau_c = 0.20\%$ in the conservative scenario. A description of this approach can be found in Section IV of the Internet Appendix. (Color figure can be viewed at wileyonlinelibrary.com)

factors may have worse average performance if poor performance makes publication more difficult or less desirable. Alternatively, unobserved factors could have strong performance if people chose to trade on them in secret rather than publish them. Either way, we next show how unobserved factors can be addressed in our framework.

The key insight is that the performance of factors across the universe of observed and unobserved factors is captured in our prior parameters τ_c, τ_w. Indeed, large values of these priors correspond to a large dispersion of alphas (i.e., a lot of large alphas "out there"), while small values mean that most true alphas are close to zero. Therefore, a smaller τ leads to stronger shrinkage toward zero for our posterior alphas, leading to fewer factor "discoveries" and a lower replication rate. Figure 10 shows how our estimated replication rate depends on the most important prior parameter, τ_c, based on the τ_w that we estimate from the data.[42]

In Figure 10, we show how the replication rate varies with τ_c in precise quantitative terms. Note that while the replication rate does indeed rise with τ_c, the differences are small in magnitude across a large range of τ_c values, demonstrating robustness of our conclusions about replicability.

[42] Figure IA.6 shows that the results are robust to alternative values of τ_w.

The stable replication rate in Figure 10 also suggests that the replication rate among the observed factors would be similar even if we had observed the unobserved factors. The figure highlights several key values of τ_c: both the value of τ_c that we estimated from the observed data (as explained in Appendix B) and values that adjust for unobserved data in different ways.

We adjust τ_c for unobserved factors as follows. We simulate a data set that proxies for the full set of factors in the population (including those that are unobserved), and then estimate the τ's that match this sample. One set of simulations is constructed to match the baseline scenario of Harvey, Liu, and Zhu (2016, table 5.A, row 1), which estimates that researchers have tried $M = 1,297$ factors, of which 39.6% have zero alpha and the rest have a Sharpe ratio of 0.44. We also consider the more conservative scenario of Harvey, Liu, and Zhu (2016, table 5.B, row 1), which implies that researchers have tried $M = 2,458$ factors, of which 68.3% have zero alpha. Section IV of the Internet Appendix provides more details on these simulations. The result, as seen in Figure 10, is that values of τ_c that correspond to these scenarios from Harvey, Liu, and Zhu (2016) still lead to a conclusion of a high replication rate in our factor universe. The replication rate is 81.5%, and 79.8% for the prior hyperparameters implied by the baseline and conservative scenario, respectively.

A closely related bias is that factors may suffer from alpha-hacking as discussed in Section I.A (Proposition 1), which makes realized in-sample factor returns too high. To account for this bias, we estimate the prior hyperparameters using only out-of-sample data. The estimated values are $\tau_c = 0.27\%$ and $\tau_w = 0.22\%$. These hyperparameters are similar to those implied by the baseline scenario of Harvey, Liu, and Zhu (2016) as seen in Figure 10. With these hyperparameters, the replication rate is 81.5%.

D. Economic Significance of Factors

Which factors (and which themes) are the most impactful anomalies in economic terms? We shed light on this question by identifying which factors matter most from an investment performance standpoint.

Figure 11 shows the alpha confidence intervals for all world factors, sorted by the median posterior alpha within clusters. This figure is similar to Figure 4, but now we focus on the world instead of the U.S. factors, and here we sort factors into clusters. We also focus on factors that the original studies conclude are significant. We see that world factor alphas tend to be economically large, often above 0.3% per month, and highly significant in most clusters. The exception is the low-leverage cluster, where we also see a low replication rate in preceding analyses.

D.1. By Region and By Size

We next consider which factors are most economically important across global regions and across stock size groups. In Panel A of Figure 12, we construct factors using only stocks in the five size subsamples presented earlier

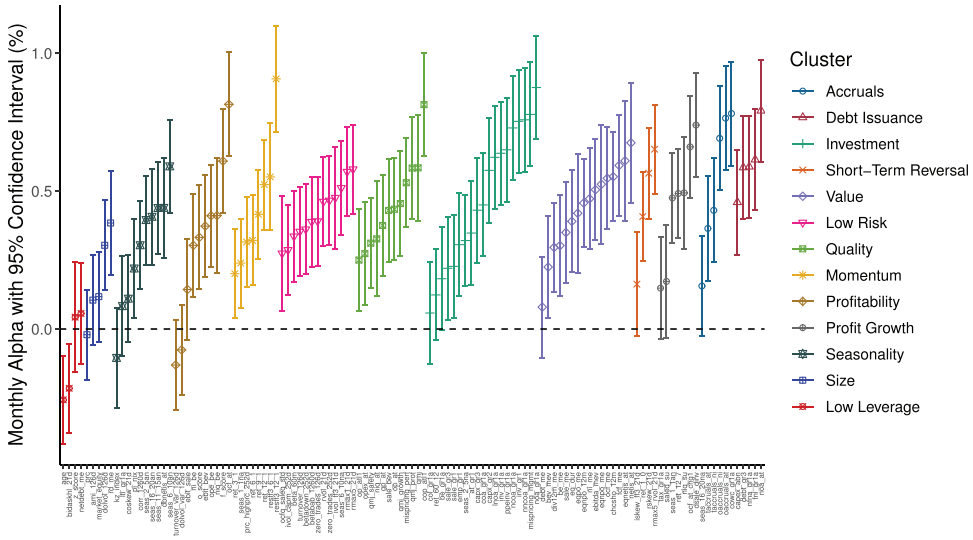

Figure 11. World alpha posterior by factor and cluster. The figure reports the EB posterior 95% confidence interval for the true alpha of a world factor create as a capitalization-weighted average of all country-specific factors in our data set. We only include factors that the original paper finds significant. (Color figure can be viewed at wileyonlinelibrary.com)

in Figure 5, namely, mega, large, small, micro, and nano stock samples. For each sample, we calculate cluster-level alphas as the equal-weighted average alpha of rank-weighted factors within the cluster.[43] We see, perhaps surprisingly, that the ordering and magnitude of alphas is broadly similar across size groups. The Spearman rank correlation of alphas for mega caps versus micro caps is 73%. Only the nano stock sample, defined as stocks below the 1st percentile of the NYSE size distribution (which amounted to 458 out of 4,356 stocks in the United States at the end of 2020), exhibits notable deviation from the other groups. The Spearman rank correlation between alphas of mega caps and nano caps is 36%.

Panel B of Figure 12 shows cluster-level alphas across regions. Again, we find consistency in alphas across the globe, with the obvious standout being the size theme, which is much more important in emerging markets than in developed markets. U.S. factor alphas share a 62% Spearman correlation with the developed ex.-U.S. sample, and a 43% correlation with the emerging markets sample.

D.2. Controlling for Other Themes

We have focused so far on whether factors (or clusters) possess significant positive alpha relative to the market. The limitation of studying factors in

[43] Rank-weighting is similar to equal-weighting and used here to illustrate the performance of typical stocks in each size group. See equation (1) in Asness, Moskowitz, and Pedersen (2013).

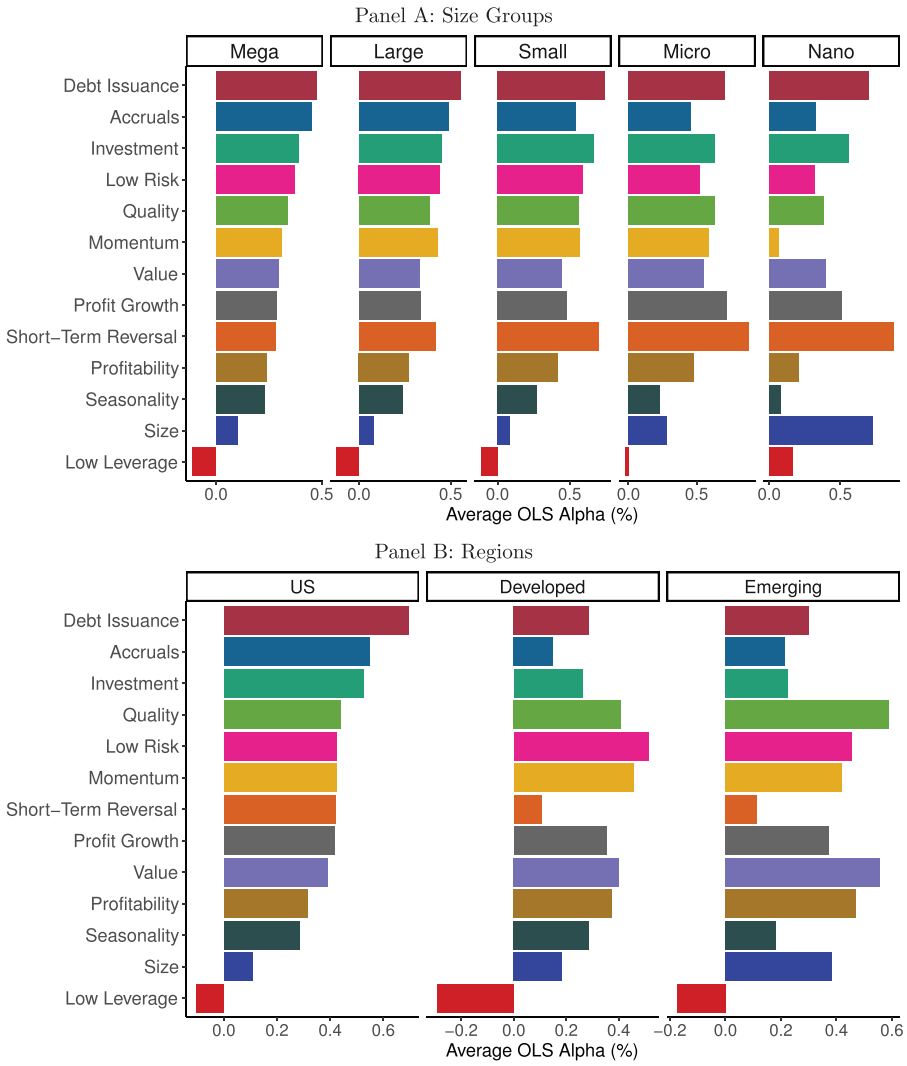

Figure 12. Alphas by stock size group and geographic region. The figure presents average cluster-level alphas for factors formed from subsamples defined by different stock market capitalization groups (Panel A) and regions (Panel B). (Color figure can be viewed at wileyonlinelibrary.com)

terms of CAPM alpha is that it does not control for duplicate behavior other than through the market factor. Economically important factors are those that have large impact on an investor's overall portfolio, and this requires understanding which clusters contribute alpha while controlling for all others.

To this end, we estimate cluster weights in a tangency portfolio that invests jointly in all cluster-level portfolios. We test the significance of the estimated weights using the method of Britten-Jones (1999). In addition to

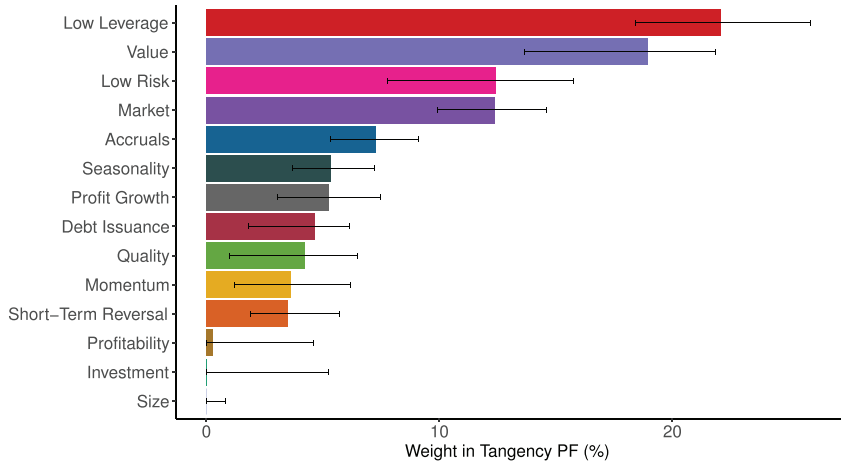

Figure 13. Tangency portfolio weights. The returns are from U.S. portfolios. We compute the cluster return as the equal-weighted return of all factors with data available at a given point in time. We further add the U.S. market return. We estimate the tangency weights following the method of Britten-Jones (1999) with a nonnegativity constraint. The error bars are 90% confidence intervals based on 10,000 bootstrap samples and the percentile method. The data start in 1952 to ensure that all clusters have nonmissing observations. (Color figure can be viewed at wileyonlinelibrary.com)

our 13 cluster-level factors, we also include the market portfolio as a way to benchmark factors to the CAPM null. Lastly, we constrain all weights to be nonnegative (because we have signed the factors to have positive expected returns according to the findings of the original studies).

Figure 13 reports the estimated tangency portfolio weights and their 90% bootstrap confidence intervals. When a factor has a significant weight in the tangency portfolio, it means that it matters for an investor, even controlling for all the other factors. We see that all but three clusters are significant in this sense. We also see that conclusions about cluster importance change when clusters are studied jointly. For example, value factors become stronger when controlling for other effects because of their hedging benefits relative to momentum, quality, and low leverage. More surprisingly, the low-leverage cluster becomes one of the most heavily weighted clusters, in large part due to its ability to hedge value and low-risk factors. The hedging performance of value and low-leverage clusters is clearly discernible in Table IA.16, which shows the average pairwise correlations among factors within and across clusters.[44] Section VI of the Internet Appendix provides further performance attribution of the tangency portfolio at the factor level.[45]

[44] Tables IA.9 and IA.10 show how tangency portfolio weights vary by region and by size group.

[45] Figure IA.11 shows the performance of each cluster together with the market portfolio, Figure IA.12 shows how the optimal portfolio changes when one cluster is excluded, and Figure IA.14 shows the importance of each factor for the optimal portfolio.

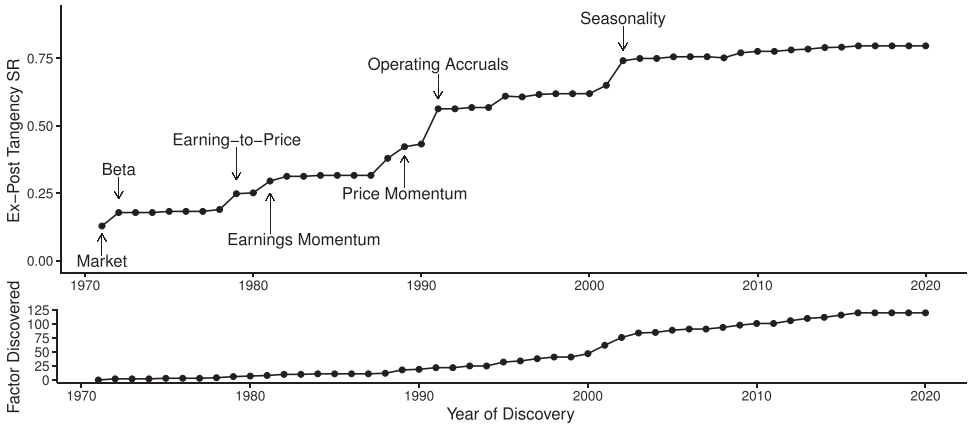

Figure 14. The evolution of the tangency Sharpe ratio. The top panel shows the Sharpe ratio on the ex post tangency portfolio. A factor is included in the tangency portfolio only after the end of the sample in which the factor was studied in the original publication (we only include factors that were found to be significant in the original paper). We highlight selected factors that significantly improve the optimal portfolio, starting with the market portfolio. We use the longest available balanced U.S. sample, 1972 to 2020 (i.e., when all factors are available).

D.3. Evolution of Finance Factor Research

The number of published factors has increased over time as seen in the bottom panel of Figure 14. To what extent have these new factors continued to add new insights versus simply repackage existing information?

To address this question, we consider how the optimal risk-return tradeoff has evolved over time as factors have been discovered. Specifically, Figure 14 computes the monthly Sharpe ratio of the ex post tangency portfolio that only invests in factors discovered by a certain point in time.[46] The starting point (on the left) of the analysis is the 0.13 Sharpe ratio of the market portfolio in the U.S. sample over 1972 to 2020 when all factors are available. The ending point (on the right) is the 0.80 Sharpe ratio of the tangency portfolio that invests the optimal weights across all factors over the same U.S. sample period.[47] In between, we see how the Sharpe ratio of the tangency portfolio has evolved as factors have been discovered. The improvement is gradual over time, but we also see occasional large increases when researchers have discovered especially impactful factors (usually corresponding to new themes in our classification scheme). An example is the operating accruals factor proposed by Sloan (1996),

[46] We estimate tangency portfolio weights following the method of Pedersen, Babu, and Levine (2021), which offers a sensible approach to mean-variance optimization for high-dimensional data. Estimation details are provided in Section VI of the Internet Appendix.

[47] The high Sharpe ratio partly reflects the fact that we are conducting in-sample optimization. If we instead run a pseudo out-of-sample analysis via cross-validation, we find a monthly Sharpe ratio of 0.56.

which increased the tangency Sharpe ratio from 0.43 to 0.56. More recently, the seasonality factors of Heston and Sadka (2008) have increased the Sharpe ratio from 0.65 to 0.74.

IV. Conclusion: Finance Research Posterior

We introduce a hierarchical Bayesian model of alphas that emphasizes the joint behavior of factors and provides a more powerful MT adjustment than common frequentist methods. Based on this framework, we revisit the evidence on replicability in factor research and come to substantially different conclusions than prior literature. We find that U.S. equity factors have a high degree of internal validity in the sense that over 80% of factors remain significant after modifications in factor construction that make all factors consistent and more implementable while still capturing the original signal (Hamermesh (2007)) and after accounting for MT concerns (Harvey, Liu, and Zhu (2016), Harvey (2017)).

We also provide new evidence demonstrating a high degree of external validity in factor research. In particular, we find highly similar qualitative and quantitative behavior in a large sample of 153 factors across 93 countries as we find in the United States. We also show that, within the United States, factors exhibit a high degree of consistency between their published in-sample results and out-of-sample data not considered in the original studies. We show that some out-of-sample factor decay is to be expected in light of Bayesian posteriors based on publication evidence. Therefore, the new evidence from postpublication data largely confirms the Bayesian's beliefs, which has led to relatively stable Bayesian alpha estimates over time.

In addition to providing a powerful tool for replication, our Bayesian framework has several additional applications. For example, the model can be used to correctly interpret out-of-sample evidence, look for evidence of alpha-hacking, compute the expected number of false discoveries and other relevant statistics based on the posterior, analyze portfolio choice taking into account both estimation uncertainty and return volatility, and evaluate asset pricing models.

Finally, the code, data, and meticulous documentation for our analysis are available online. Our large global factor data set and the underlying stock-level characteristics are easily accessible to researchers by using our publicly available code and its direct link to WRDS. Our database will be updated regularly with new data releases and code improvements. We hope that our methodology and data will help promote credible finance research.

Initial submission: February 1, 2021; Accepted: July 16, 2021
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix A: Additional Results and Proofs

A.1. Additional Results on Alpha-Hacking

We consider the situation where the researcher has in-sample data from time 1 to time T and an out-of-sample (oos) period from time $T+1$ to $T+T^{oos}$. The researcher may have used alpha-hacking during the in-sample period, but this does not affect the out-of-sample period. The researcher is interested in the posterior alpha based on the total evidence, in-sample and out-of-sample, which is useful for predicting factor performance in a future time period (i.e., a time period that is out-of-sample relative to the existing out-of-sample period).

PROPOSITION A.1 (Out-of-sample alpha): *The posterior alpha based on in-sample data from time 1 to T with alpha-hacking, and an out-of-sample period from $T+1$ to $T+T^{oos}$ is given by*

$$E(\alpha|\hat{\alpha},\hat{\alpha}^{oos}) = \kappa^{oos}\big(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\big), \tag{A1}$$

where $w=\frac{\sigma^2/T^{oos}}{\bar{\sigma}^2/T+\sigma^2/T^{oos}}\in(0,1)$ is the relative weight on the in-sample period relative to the out-of-sample period and $\kappa^{oos}=\frac{1}{1+1/(\tau^2([\bar{\sigma}^2/T]^{-1}+[\sigma^2/T^{oos}]^{-1}))}$ is a shrinkage parameter.

We see that the more alpha-hacking the researcher has done (higher $\bar{\sigma}$), the less weight we put on the in-sample period relative to the out-of-sample period. Further, the in-sample period has the nonproportional discounting due to alpha-hacking ($\bar{\varepsilon}$), which we do not have for out-of-sample evidence.

This result formalizes the idea that an in-sample backtest plus live performance is *not* the same as a longer backtest. For example, 10 years of backtest plus 10 years of live performance is more meaningful that 20 years of backtest with no live performance. The difference is that the out-of-sample performance is free from alpha-hacking.

A.2. Proofs and Lemmas

The proofs make repeated use of the following well-known property of multivariate Normally distributed random variable. If x and y are multivariate Normal:

$$\begin{bmatrix}x\\y\end{bmatrix}\sim N\left(\begin{bmatrix}\mu_x\\\mu_y\end{bmatrix},\begin{bmatrix}\Sigma_{xx}&\Sigma_{yx}\\\Sigma_{xy}&\Sigma_{yy}\end{bmatrix}\right), \tag{A2}$$

then the conditional distribution of x given y has the following Normal distribution:

$$x|y\sim N\big(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y),\ \Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\big). \tag{A3}$$

The proofs also make use of the following two lemmas.

LEMMA A.1: *For random variables x, y, and z, it holds that $E(Var(x|y,z)) \leq E(Var(x|y))$ and, if the random variables are jointly normal, then $Var(x|y,z) \leq Var(x|y)$.*

LEMMA A.2: *Let A be an $N \times N$ matrix for which all diagonal elements equal a and all off-diagonal elements equal b, where $a \neq b$ and $a + b(N-1) \neq 0$. Then the inverse A^{-1} exists and is of the same form:*

$$A = \begin{bmatrix} a & & b \\ & \ddots & \\ b & & a \end{bmatrix} \qquad A^{-1} = \begin{bmatrix} c & & d \\ & \ddots & \\ d & & c \end{bmatrix}, \tag{A4}$$

where $c = \frac{a+b(N-2)}{(a-b)(a+b(N-1))}$ and $d = \frac{-b}{(a-b)(a+b(N-1))}$.

PROOF OF LEMMA A.1: Using the definition of conditional variance, we have

$$E(\mathrm{Var}(x|y,z)) = E(E(x^2|y,z)) - E(\big[E(x|y,z)\big]^2) = E(x^2) - E(\big[E(x|y,z)\big]^2).$$

Hence, using Jensen's inequality, we have

$$\begin{aligned} E(\mathrm{Var}(x|y)) - E(\mathrm{Var}(x|y,z)) =& E(\big[E(x|y,z)\big]^2) - E(\big[E(x|y)\big]^2) \\ =& E(\big[E(x|y,z)\big]^2) - E(\big[E(E(x|y,z)|y)\big]^2) \\ \geq& E(\big[E(x|y,z)\big]^2) - E(E(\big[E(x|y,z)\big]^2|y)) = 0. \end{aligned}$$

The result for normal distributions follows from the fact that normal conditional variances are nonstochastic, that is, $\mathrm{Var}(x|y) = E(\mathrm{Var}(x|y))$. In this case, we can also characterize the extra drop in variance due to conditioning on z using its orthogonal component ε from the regression $z = a + by + \varepsilon$, using similar notation as (A2):

$$\begin{aligned} \mathrm{Var}(x|y,z) = \mathrm{Var}(x|y,\varepsilon) =& \Sigma_{x,x} - \Sigma_{x,(y,\varepsilon)}\Sigma^{-1}_{(y,\varepsilon),(y,\varepsilon)}\Sigma_{(y,\varepsilon),x} \\ =& \Sigma_{x,x} - \Sigma_{x,y}\Sigma^{-1}_{y,y}\Sigma_{y,x} - \Sigma_{x,\varepsilon}\Sigma^{-1}_{\varepsilon,\varepsilon}\Sigma_{\varepsilon,x} = \mathrm{Var}(x|y) - \Sigma_{x,\varepsilon}\Sigma^{-1}_{\varepsilon,\varepsilon}\Sigma_{\varepsilon,x}. \end{aligned}$$

□

PROOF OF LEMMA A.2: The proof follows from inspection: The product of A and its proposed inverse clearly has the same form as A with diagonal elements

$$ac + bd(I-1) = \frac{a(a+b(N-2)) - b^2(N-1)}{(a-b)(a+b(N-1))} = \frac{a^2 + ab(N-1) - ab - b^2(N-1)}{(a-b)(a+b(N-1))} = 1$$

and off-diagonal elements

$$ad + bc + bd(N-2) = \frac{-ab + b(a+b(N-2)) - b^2(N-2)}{(a-b)^2(a+b(N-1))^2} = 0.$$

In other words, AA^{-1} equals the identity, proving the result. □

PROOF OF EQUATIONS (4) to (6): The posterior distribution of the true alpha given the observed factor return is computed using (A3). The conditional mean is

$$E(\alpha|\hat{\alpha}) = 0 + \frac{\text{Cov}(\alpha, \hat{\alpha})}{\text{Var}(\hat{\alpha})}(\hat{\alpha} - 0) = \frac{\tau^2}{\tau^2 + \sigma^2/T}\hat{\alpha} = \kappa\hat{\alpha},$$

where κ is given by (5) and the posterior variance is

$$\text{Var}(\alpha|\hat{\alpha}) = \text{Var}(\alpha) - \frac{(\text{Cov}(\alpha, \hat{\alpha}))^2}{\text{Var}(\hat{\alpha})} = \tau^2 - \tau^2\frac{\tau^2}{\tau^2 + \sigma^2/T} = \frac{\tau^2\sigma^2/T}{\tau^2 + \sigma^2/T} = \kappa\frac{\sigma^2}{T}.$$

□

PROOF OF PROPOSITION 1: The posterior alpha with alpha-hacking is given via (A3) as

$$E(\alpha|\hat{\alpha}) = 0 + \frac{\text{Cov}(\alpha, \hat{\alpha})}{\text{Var}(\hat{\alpha})}(\hat{\alpha} - E(\hat{\alpha})) = \frac{\tau^2}{\tau^2 + \bar{\sigma}^2/T}(\hat{\alpha} - \bar{\varepsilon}) = -\kappa_0 + \kappa^{\text{hacking}}\hat{\alpha},$$

where $\kappa^{\text{hacking}} = \frac{1}{1+\frac{\bar{\sigma}^2}{\tau^2 T}}$, $\kappa_0 = \kappa^{\text{hacking}}\bar{\varepsilon} \geq 0$, and $\kappa^{\text{hacking}} \leq \kappa$ because $\bar{\sigma} \geq \sigma$. □

PROOF OF PROPOSITION 2: The posterior mean given $\hat{\alpha}$ and $\hat{\alpha}^g$ is computed via (A3) as

$$\begin{aligned}
E(\alpha|\hat{\alpha}, \hat{\alpha}^g) &= \begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2 + \sigma_T^2 & \tau^2 + \rho\sigma_T^2 \\ \tau^2 + \rho\sigma_T^2 & \tau^2 + \sigma_T^2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\alpha} \\ \hat{\alpha}^g\end{bmatrix} \\
&= \frac{1}{\det}\begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2 + \sigma_T^2 & -(\tau^2 + \rho\sigma_T^2) \\ -(\tau^2 + \rho\sigma_T^2) & \tau^2 + \sigma_T^2\end{bmatrix}\begin{bmatrix}\hat{\alpha} \\ \hat{\alpha}^g\end{bmatrix} \\
&= \frac{\tau^2(1-\rho)\sigma_T^2}{\det}\left(\hat{\alpha} + \hat{\alpha}^g\right) \\
&= \frac{\tau^2(1-\rho)}{\sigma_T^2(1-\rho)(1+\rho) + 2\tau^2(1-\rho)}\left(\hat{\alpha} + \hat{\alpha}^g\right) \\
&= \kappa^g\left(\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g\right)
\end{aligned}$$

using the notation $\sigma_T^2 = \sigma^2/T$ and

$$\det = (\tau^2 + \sigma_T^2)^2 - (\tau^2 + \rho\sigma_T^2)^2 = \sigma_T^2[\sigma_T^2(1-\rho^2) + 2\tau^2(1-\rho)].$$

The global shrinkage parameter κ^g is in $[\kappa, 1]$ and decreases with the correlation ρ, attaining the minimum value, $\kappa^g = \kappa$, when $\rho = 1$ as is clearly seen from (12).

The result about the posterior variance follows from Lemma A.1. □

PROOF OF PROPOSITION 3: The prior joint distribution of the true and estimated alphas is given by the following expression, where we focus on factor 1

without loss of generality:

$$\begin{bmatrix} \alpha^1 \\ \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_c^2+\tau_w^2 & \tau_c^2+\tau_w^2 & \tau_c^2 & \cdots & \tau_c^2 \\ \tau_c^2+\tau_w^2 & \tau_c^2+\tau_w^2+\sigma^2/T & & & \tau_c^2+\rho\sigma^2/T \\ \tau_c^2 & & & & \\ \vdots & & & \ddots & \\ \tau_c^2 & \tau_c^2+\rho\sigma^2/T & & & \tau_c^2+\tau_w^2+\sigma^2/T \end{bmatrix} \right). \tag{A5}$$

The posterior alpha of factor 1 is therefore normally distributed with a mean derived using the standard formula for conditional normal distributions (A3):

$$E(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) = \begin{bmatrix} \tau_c^2+\tau_w^2 \\ \tau_c^2 \\ \vdots \\ \tau_c^2 \end{bmatrix}^\top \begin{bmatrix} \tau_c^2+\tau_w^2+\sigma^2/T & & \tau_c^2+\rho\sigma^2/T \\ & \ddots & \\ \tau_c^2+\rho\sigma^2/T & & \tau_c^2+\tau_w^2+\sigma^2/T \end{bmatrix}^{-1} \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix}.$$

We next use Lemma A.2 and its notation, that is, $a = \tau_c^2+\tau_w^2+\sigma^2/T$, $b = \tau_c^2+\rho\sigma^2/T$, and c', d are defined accordingly, where we use the notation c' to avoid confusion with the c in equation (14). This application of Lemma A.2 yields

$$\begin{aligned} E(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) &= \begin{bmatrix} \tau_c^2+\tau_w^2 \\ \tau_c^2 \\ \vdots \\ \tau_c^2 \end{bmatrix}^\top \begin{bmatrix} c' & & d \\ & \ddots & \\ d & & c' \end{bmatrix} \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \\ &= \begin{bmatrix} \tau_c^2(c'+d(N-1))+\tau_w^2 c' \\ \tau_c^2(c'+d(N-1))+\tau_w^2 d \\ \vdots \\ \tau_c^2(c'+d(N-1))+\tau_w^2 d \end{bmatrix}^\top \begin{bmatrix} \hat{\alpha}^1 \\ \vdots \\ \hat{\alpha}^N \end{bmatrix} \\ &= (\tau_c^2(c'+d(N-1))+\tau_w^2 d)N\hat{\alpha}^\cdot + \tau_w^2(c'-d)\hat{\alpha}^1 \\ &= (\tau_c^2\frac{N}{a+b(N-1)} - \tau_w^2\frac{bN}{(a-b)(a+b(N-1))})\hat{\alpha}^\cdot + \tau_w^2\frac{1}{a-b}\hat{\alpha}^1 \\ &= \frac{\tau_c^2}{b+\frac{a-b}{N}}\hat{\alpha}^\cdot + \frac{\tau_w^2}{a-b}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{a-b}{bN}}\hat{\alpha}^\cdot\right) \\ &= \frac{\tau_c^2}{\tau_c^2+\rho\sigma^2/T+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{N}}\hat{\alpha}^\cdot + \frac{\tau_w^2}{\tau_w^2+(1-\rho)\sigma^2/T}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\hat{\alpha}^\cdot\right) \\ &= \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{\tau_c^2 N}}\hat{\alpha}^\cdot + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^1 - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\hat{\alpha}^\cdot\right). \end{aligned}$$

The posterior has conditional variance

$$\begin{aligned}
\mathrm{Var}(\alpha^1|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) &= \tau_c^2+\tau_w^2-\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}^\top\begin{bmatrix}c' & & d\\ & \ddots & \\ d & & c'\end{bmatrix}\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}\\
&= \tau_c^2+\tau_w^2-\begin{bmatrix}\tau_c^2(c'+d(N-1))+\tau_w^2c'\\ \tau_c^2(c'+d(N-1))+\tau_w^2d\\ \vdots\\ \tau_c^2(c'+d(N-1))+\tau_w^2d\end{bmatrix}^\top\begin{bmatrix}\tau_c^2+\tau_w^2\\ \tau_c^2\\ \vdots\\ \tau_c^2\end{bmatrix}\\
&= \tau_c^2+\tau_w^2-(\tau_c^2(c'+d(N-1))+\tau_w^2c')(\tau_c^2+\tau_w^2)\\
&\quad -(\tau_c^2(c'+d(N-1))+\tau_w^2d)\tau_c^2(N-1)\\
\to\quad &\tau_c^2+\tau_w^2-(\tau_c^2(\frac{1}{a-b}-\frac{1}{a-b})+\tau_w^2\frac{1}{a-b})(\tau_c^2+\tau_w^2)\\
&\quad -(\tau_c^2\frac{1}{b}-\tau_w^2\frac{1}{a-b})\tau_c^2\\
&= \tau_c^2+\tau_w^2-\left(\tau_w^4\frac{1}{a-b}+\tau_c^4\frac{1}{b}\right)\\
&= \tau_c^2+\tau_w^2-\left(\frac{\tau_w^4}{\tau_w^2+(1-\rho)\sigma^2/T}+\frac{\tau_c^4}{\tau_c^2+\rho\sigma^2/T}\right).
\end{aligned}$$

The last results follow from Lemma A.1. □

PROOF OF PROPOSITION 4: We write the joint prior distribution of true and observed alphas in the multilevel hierarchical model as

$$\begin{pmatrix}\alpha\\ \hat{\alpha}\end{pmatrix}\sim N\left(\alpha^0\,1_{2NK},\begin{pmatrix}\Omega & \Omega\\ \Omega & \Omega+\Sigma/T\end{pmatrix}\right). \tag{A6}$$

The posterior mean vector of true alphas is computed via (A3):

$$\begin{aligned}
E(\alpha|\hat{\alpha}) &= 1_{NK}\alpha_0+\Omega(\Omega+\Sigma/T)^{-1}(\hat{\alpha}-1_{NK}\alpha_0)\\
&= \left(\Omega^{-1}+T\Sigma^{-1}\right)^{-1}\left(\Omega^{-1}1_{NK}\alpha_0+T\Sigma^{-1}\hat{\alpha}\right),
\end{aligned}$$

using the fact that $(\Omega+\Sigma/T)^{-1}=\Omega^{-1}-\Omega^{-1}(\Omega^{-1}+T\Sigma^{-1})^{-1}\Omega^{-1}$ by the Woodbury matrix identity. The posterior variance is computed similarly via (A3) and the same application of the Woodbury matrix identity as

$$\mathrm{Var}(\alpha|\hat{\alpha})=\Omega-\Omega(\Omega+\Sigma/T)^{-1}\Omega=\left(\Omega^{-1}+T\Sigma^{-1}\right)^{-1}.$$

□

PROOF OF PROPOSITION 5: Based on the definition of the Bayesian FDR, we have:

$$\begin{aligned}
\text{FDR}^{\text{Bayes}} &= E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}}\middle|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau\right)\\
&= \frac{1}{\sum_i 1_{\{i \text{ discovery}\}}} E\left(\sum_i 1_{\{i \text{ false discovery}\}}\middle|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau\right)\\
&= \frac{1}{\sum_i 1_{\{i \text{ discovery}\}}} \sum_i Pr(i \text{ false discovery}|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau)\\
&= \frac{1}{\#\text{discoveries}} \sum_{i \text{ discovery}} p\text{-null}_i\\
&\leq 2.5\%.
\end{aligned}$$

□

PROOF OF PROPOSITION A.1: The posterior mean alpha is

$$\begin{aligned}
E(\alpha|\hat{\alpha},\hat{\alpha}^{oos}) &= \begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\bar{\sigma}_T^2 & \tau^2\\ \tau^2 & \tau^2+\sigma_{oos}^2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\alpha}-\bar{\varepsilon}\\ \hat{\alpha}^{oos}\end{bmatrix}\\
&= \frac{1}{\det}\begin{bmatrix}\tau^2 & \tau^2\end{bmatrix}\begin{bmatrix}\tau^2+\sigma_{oos}^2 & -\tau^2\\ -\tau^2 & \tau^2+\bar{\sigma}_T^2\end{bmatrix}\begin{bmatrix}\hat{\alpha}-\bar{\varepsilon}\\ \hat{\alpha}^{oos}\end{bmatrix}\\
&= \frac{\tau^2}{\det}\left(\sigma_{oos}^2(\hat{\alpha}-\bar{\varepsilon})+\bar{\sigma}_T^2\hat{\alpha}^g\right)\\
&= \frac{\tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)}{\tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)+\bar{\sigma}_T^2\sigma_{oos}^2}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right)\\
&= \frac{\tau^2}{\tau^2+\bar{\sigma}_T^2\sigma_{oos}^2/(\bar{\sigma}_T^2+\sigma_{oos}^2)}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right)\\
&= \frac{1}{1+\frac{1}{\tau^2(\bar{\sigma}_T^{-2}+\sigma_{oos}^{-2})}}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right),
\end{aligned}$$

using the notation $\bar{\sigma}_T^2=\bar{\sigma}^2/T$, $\sigma_{oos}^2=\sigma^2/T^{oos}$, and

$$\det=(\tau^2+\bar{\sigma}_T^2)(\tau^2+\sigma_{oos}^2)-\tau^4=\tau^2(\bar{\sigma}_T^2+\sigma_{oos}^2)+\bar{\sigma}_T^2\sigma_{oos}^2.$$

□

Appendix B: EB Estimation

For convenient reference, we restate the multilevel hierarchical model of Section I. For factor i in cluster j corresponding to signal n, the factor is

$$f_t^i = \alpha^i + \beta^i r_t^m + \varepsilon_t^i$$

with

$$\alpha^i = \alpha^o + c^j + s^n + w^i,$$

where the alpha components are $\alpha^o = 0$, $c^j \sim N(0, \tau_c^2)$, $s^n \sim N(0, \tau_s^2)$, and $w^i \sim N(0, \tau_w^2)$. We write alpha in vector form as

$$\alpha = \alpha^o \, 1_{NK} + Mc + Zs + w, \tag{B1}$$

where $\alpha = (\alpha^1, \ldots, \alpha^{NK})'$, $c = (c^1, \ldots, c^J)'$, $s = (s^1, \ldots, s^N)'$, $w = (w^1, \ldots, w^{NK})'$, M is the $NK \times J$ matrix of cluster memberships, and Z is the $NK \times N$ matrix indicating the characteristic that factor i is based on. Given the hyperparameters $(\alpha^0, \tau_c, \tau_s, \tau_w)$, the prior mean and covariance matrix of alphas are

$$E[\alpha] = 0, \quad \Omega \equiv \mathrm{Var}(\alpha) = MM'\tau_c^2 + ZZ'\tau_s^2 + I_{NK}\tau_w^2. \tag{B2}$$

The vector of return shocks is $\varepsilon_t = (\varepsilon_t^1, \ldots, \varepsilon_t^{NK})'$, which is distributed $\varepsilon_t \sim N(0, \Sigma)$.

Given this structure, we estimate the model as follows. The vector of factor returns $f_t = (f_t^1, \ldots, f_t^{NK})'$ has marginal likelihood—that is, after integrating out the uncertain alpha components—that is distributed as

$$f_t \sim N(0, [\Omega + \Sigma]),$$

or, equivalently (treating CAPM betas as known), the estimated alphas are distributed[48]

$$\hat{\alpha} \sim N(0, [\Omega + \Sigma/T]).$$

The matrices Z and M are given by the factor definition and cluster assignment (Table IA.III), respectively. We use a plug-in estimate of the factor CAPM-residual return covariance matrix, denoted $\hat{\Sigma}$ (discussed below). Finally, given $\hat{\Sigma}$, Z, and M, we estimate the hyperparameters of the prior distribution, (τ_c, τ_s, τ_w), via MLE based on the marginal likelihood.

This estimation approach is an example of the EB method. It approximates the fully Bayesian posterior calculation (which requires integrating over a hyperprior distribution of hyperparameters, usually an onerous calculation) by setting the hyperparameters to their most likely values based on the marginal likelihood. It is particularly well suited to hierarchical Bayesian models in which parameters for individual observations share some common structure,

[48] We abstract from uncertainty in CAPM betas to emphasize the Bayesian updating of alphas. Our conclusions are qualitatively insensitive to accounting for beta uncertainty.

so that the realized heterogeneity across individual observations is informative about sensible values for the hyperparameters of the prior. Our model and estimation approach implementation is a minor variation on Bayesian hierarchical normal mean models that are common in Bayesian statistics (textbook treatments include Efron (2012), Gelman et al. (2013), and Maritz (2018)). We conduct sensitivity analysis to ensure that our results are robust to a wide range of hyperparameters (see Figure IA.6). Also, we note that our EB methodology is more easily replicable than a full-Bayesian setting with additional hyperpriors as EB relies on closed-form Bayesian updating rather than numerical integration.

To ensure cross-sectional stationarity, we scale each factor such that their monthly idiosyncratic volatility is $10\%/\sqrt{12}$ (i.e., 10% annualized). To construct a plug-in estimate of the factor residual return covariance matrix, denoted $\hat{\Sigma}$, we face two main empirical challenges. First, the sample covariance is poorly behaved due the relatively large number of factors compared to the number of time-series observations. Second, we have an unbalanced panel because different factors come online at different points in time. To address the first challenge, we impose a block equicorrelation structure on Σ based on factors' cluster membership.[49] The correlation between factors in clusters i and j is estimated as the average correlation among all pairs such that one factor is in cluster i and the other is in j. In our global analyses, blocks correspond to region-cluster pairs. To address unbalancedness, we use the bootstrap. In particular, we generate 10,000 bootstrap samples that resample rows of the unbalanced factor return data set. Each bootstrap sample is, therefore, also unbalanced, and we use this to produce a distribution of alpha estimates. From this we calculate $\hat{\Sigma}/T$ as the covariance of alphas across bootstrap samples (imposing the block equicorrelation structure).

Table I shows the estimated hyperparameters across different samples. While most of our analysis is based on these full-sample estimates, we also consider rolling estimates of the hyperparameters when considering out-of-sample evidence as seen in Figure IA.5.

REFERENCES

Acharya, Viral, and Lasse Heje Pedersen, 2005, Asset pricing with liquidity risk, *Journal of Financial Economics* 77, 375–410.

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, *Journal of Financial Economics* 17, 223–249.

Asness, Cliff, Tobias Moskowitz, and Lasse Heje Pedersen, 2013, Value and momentum everywhere, *Journal of Finance* 68, 929–985.

[49] As advocated by Engle and Kelly (2012) and Elton, Gruber, and Spitzer (2006), block equicorrelation constrains all pairs of factors in the same block to share a single correlation parameter, and likewise for cross-block correlations. This stabilizes covariance matrix estimates by dramatically reducing the parameterization of the correlation matrix, while leaving the individual variance estimates unchanged.

Asness, Clifford, and Andrea Frazzini, 2013, The devil in HML's details, *Journal of Portfolio Management* 39, 49–68.

Benjamini, Yoav, and Yosef 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, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29, 1165–1188.

Berry, Donald A., and Yosef Hochberg, 1999, Bayesian perspectives on multiple comparisons, *Journal of Statistical Planning and Inference* 82, 215–227.

Bettis, Richard A, 2012, The search for asterisks: Compromised statistical tests and flawed theories, *Strategic Management Journal* 33, 108–113.

Britten-Jones, Mark, 1999, The sampling error in estimates of mean-variance efficient portfolio weights, *Journal of Finance* 54, 655–671.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2023, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *Journal of Finance*, 78, 487–557.

Chen, Andrew Y., 2021, The limits of p-hacking: Some thought experiments, *Journal of Finance*, 76, 2447–2480.

Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review*, 11, 207–264.

Chen, Andrew Y., and Tom Zimmermann, 2020b, Publication bias and the cross-section of stock returns, *Review of Asset Pricing Studies* 10, 249–289.

Chinco, Alex, Andreas Neuhierl, and Michael Weber, 2021, Estimating the anomaly base rate, *Journal of Financial Economics* 140, 101–126.

Chordia, Tarun, Amit Goyal, and Alessio Saretto, 2020, Anomalies and false rejections, *Review of Financial Studies* 33, 2134–2179.

Cochrane, John H., 2011, Presidential address: Discount rates, *Journal of Finance* 66, 1047–1108.

Efron, Bradley, 2007, Size, power and false discovery rates, *Annals of Statistics* 35, 1351–1377.

Efron, Bradley, 2012, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1 (Cambridge University Press, Cambridge).

Efron, Bradley, and Robert Tibshirani, 2002, Empirical Bayes methods and false discovery rates for microarrays, *Genetic Epidemiology* 23, 70–86.

Elton, Edwin J., Martin J. Gruber, and Jonathan Spitzer, 2006, Improved estimates of correlation coefficients and their impact on optimum portfolios, *European Financial Management* 12, 303–318.

Engle, Robert, and Bryan Kelly, 2012, Dynamic equicorrelation, *Journal of Business & Economic Statistics* 30, 212–228.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Gelman, Andrew, 2016, Bayesian inference completely solves the multiple comparisons problem, *Statistical Modeling, Causal Inference, and Social Science*. Available at https://statmodeling.stat.columbia.edu/2016/08/22/bayesian-inference-completely-solves-the-multiple-comparisons-problem/.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, 2013, *Bayesian Data Analysis*:, third edition (CRC Press, New York).

Gelman, Andrew, Jennifer Hill, and Masanao Yajima, 2012, Why we (usually) don't have to worry about multiple comparisons, *Journal of Research on Educational Effectiveness* 5, 189–211.

Green, Jeremiah, John R.M. Hand, and X. Frank Zhang, 2017, The characteristics that provide independent information about average us monthly stock returns, *Review of Financial Studies* 30, 4389–4436.

Greenland, Sander, and Albert Hofman, 2019, Multiple comparisons controversies are about context and costs, not frequentism versus Bayesianism, *European Journal of Epidemiology* 34, 801–808.

Greenland, Sander, and James M. Robins, 1991, Empirical-Bayes adjustments for multiple comparisons are sometimes useful, *Epidemiology*, 2, 244–251.

Hamermesh, Daniel S., 2007, Replication in economics, *Canadian Journal of Economics/Revue canadienne d'économique* 40, 715–733.

Harvey, Campbell R., 2017, Presidential address: The scientific outlook in financial economics, *Journal of Finance* 72, 1399–1440.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, …and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Heston, Steven L., and Ronnie Sadka, 2008, Seasonality in the cross-section of stock returns, *Journal of Financial Economics* 87, 418–445.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Ilmanen, Antti, Ronen Israel, Tobias J Moskowitz, Rachel Lee, and Ashwin K. Thapar, 2021, How do factor premia vary over time? A century of evidence, *Journal of Investment Management*, 19, 15–57.

Ioannidis, John PA., 2005, Why most published research findings are false, *PLoS Medicine* 2, e124.

Jacobs, Heiko, and Sebastian Müller, 2020, Anomalies across the globe: Once public, no longer existent?, *Journal of Financial Economics* 135, 213–230.

Kelly, Bryan T., Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Koijen, Ralph S.J., Tobias J. Moskowitz, Lasse Heje Pedersen, and Evert B. Vrugt, 2018, Carry, *Journal of Financial Economics* 127, 197–225.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.

Linnainmaa, Juhani T., and Michael R. Roberts, 2018, The history of the cross-section of stock returns, *Review of Financial Studies* 31, 2606–2649.

Maniadis, Zacharias, Fabio Tufano, and John A. List, 2017, To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study, *The Economic Journal* 127, F209–F235.

Maritz, Johannes S., 2018, *Empirical Bayes Methods with Applications*, second edition (CRC Press, New York).

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

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen, 2012, Time series momentum, *Journal of Financial Economics* 104, 228–250.

Murtagh, Fionn, and Pierre Legendre, 2014, Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification* 31, 274–295.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl, 2012, Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability, *Perspectives on Psychological Science* 7, 615–631.

Pedersen, Lasse Heje, Abhilash Babu, and Ari Levine, 2021, Enhanced portfolio optimization, *Financial Analysts Journal* 77, 124–151.

Shumway, Tyler, 1997, The delisting bias in CRSP data, *Journal of Finance* 52, 327–340.

Sloan, Richard G., 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review* 71, 289–315.

Ward, Joe H., 1963, Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* 58, 236–244.

Welch, Ivoi, 2019, Reproducing, extending, updating, replicating, reexamining, and reconciling, *Critical Finance Review* 8, 301–304.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.