The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXVII, NO. 3 • JUNE 2022

Luck versus Skill in the Cross Section of Mutual Fund Returns: Reexamining the Evidence

CAMPBELL R. HARVEY and YAN LIU

ABSTRACT

While Kosowski et al. (2006, *Journal of Finance* 61, 2551–2595) and Fama and French (2010, *Journal of Finance* 65, 1915–1947) both evaluate whether mutual funds outperform, their conclusions are very different. We reconcile their findings. We show that the Fama-French method suffers from an undersampling problem that leads to a failure to reject the null hypothesis of zero alpha, even when some funds generate economically large risk-adjusted returns. In contrast, Kosowski et al. substantially overreject the null hypothesis, even when all funds have a zero alpha. We present a novel bootstrapping approach that should be useful to future researchers choosing between the two approaches.

IDENTIFYING FUNDS THAT WILL "beat the market" is one of the oldest and most challenging problems in finance—with thousands of funds, some will outperform purely by luck. Influential papers by Kosowski et al. (2006) and Fama and French (2010) employ a bootstrapping approach to try to separate luck from skill but arrive at strikingly different conclusions. Kosowski et al. (2006) find that a substantial fraction of funds outperform. In contrast, Fama and French (2010) provide evidence that no advantage exists for active compared to passive management. In this paper, we seek to shed light on why the conclusions of these two studies are so diametrically opposed when both studies use similar data and a common bootstrapping approach.

While both studies use bootstrapping, their implementations are very different. The Kosowski et al. (2006) approach bootstraps the data firm by firm and requires a minimum of 60 observations. In contrast, Fama and French (2010)

Campbell R. Harvey is with Duke University and National Bureau of Economic Research. Yan Liu is with Purdue University. We thank Ken French, Stefan Nagel (Editor), and the referees for their helpful comments. We also thank Tomasz Wisniewski for his very generous help organizing the large amount of computing resources necessary for this paper. We thank Kay Jaitly for editorial assistance. We have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Yan Liu, Purdue University, Krannert Building, 403 W. State Street, West Lafayette, IN 47907-2056; e-mail: liu2746@purdue.edu.

DOI: 10.1111/jofi.13123

bootstrap the cross section of fund returns, thereby retaining the economically important correlation structure, and require a minimum of only eight observations.

We provide a number of apples-to-apples comparisons of these techniques—for example, we require the Fama-French (2010) approach to have a minimum of 60 observations. We design a simulation study in which we know the outperforming funds in advance. Our technique is related to Harvey and Liu (2020) and is designed to capture the ability of each approach to correctly identify the outperforming funds. We provide five different comparisons that we believe will be useful to future researchers seeking to choose the most powerful technique.

Our results can be summarized as follows. Comparing test size (i.e., the Type I error rate or the probability of falsely classifying a fund as an outperformer), Kosowski et al.'s (2006) approach is substantially oversized and therefore overrejects the market efficiency hypothesis, that is, the hypothesis that no fund outperforms. In contrast, the Fama and French (2010) requirement of a minimum of eight observations leads to undersampling for certain funds in the bootstrapped simulations in that the bootstrapped sample has fewer observations than the actual sample. As a result, their approach leads to a strong asymmetry in the distribution of the bootstrapped t-statistics between undersampling and oversampling (i.e., the opposite of undersampling) for funds with a short history, which makes it difficult for their test to reject the null hypothesis. As a result, their test is undersized under the null and hence lacks power to detect outperforming funds under the alternative.[1] Reconciling these two studies, we propose two adjusted Fama and French (2010) approaches that we believe will be useful for future research. The first approach simply focuses on those funds with a certain number of observations (e.g., 60 monthly observations) and is straightforward to implement. The second approach involves dropping funds with implausible t-statistics in the bootstrapped iterations. We provide guidance on what constitutes "implausible" t-statistics using our simulation approach. Both adjustments are shown to have near-optimal size and are more powerful than the original Fama and French (2010) implementation. Applying the adjusted Fama-French methods, our evidence on mutual fund outperformance lies somewhere between Kosowski et al. (2006) and Fama and French (2010).

Our paper is related to the considerable statistics literature on bootstrap-based inference, which has become popular in finance applications. Theoretically, bootstrap-based methods may deliver substantial improvement over traditional approaches based on asymptotic theories for relatively small samples (see, for example, Beran (1998), Hall (1992), Davidson and

[1] Our findings have important implications for the interpretation of many recent papers that apply the method of either Kosowski et al. (2006) or Fama and French (2010). An incomplete list of such papers includes Chen and Liang (2007), Jiang, Yao, and Yu (2007), Busse, Goyal, and Wahal (2010), Ayadi and Kryzanowski (2011), D'Agostino, McQuinn, and Whelan (2012), Cao et al. (2013), Hau and Lai (2013), Blake et al. (2013), Busse, Goyal, and Wahal (2014), Harvey and Liu (2017), Yan and Zheng (2017), and Chordia, Goyal, and Saretto (2020).

MacKinnon (1999), Horowitz (2003)).[2] Despite the strong theoretical appeal, existing Monte Carlo experiments that support bootstrap-based tests are often based on univariate tests in stylized settings. Because a generic bootstrap test that is optimal in different contexts does not exist, it is important for researchers to study the properties of a given bootstrap approach for a particular application. We conduct such an exercise for mutual fund performance evaluation that features an unbalanced panel with a large cross section, a common factor-model benchmark across funds, and potentially a nontrivial dependence structure in fund residuals in the cross section.[3]

Another recent study that analyzes Kosowski et al. (2006) and Fama and French (2010) is Huang et al. (2020, HJLP). The authors focus on the asymptotic properties of Kosowski et al. (2006) and Fama and French (2010) and propose alternative test statistics to enhance test power. Different from their paper, we focus on the empirical performance of both papers and propose enhancements based on the original test statistics proposed in Fama and French (2010). For example, while HJLP claim that Kosowski et al. (2006)'s approach has a correct asymptotic test size, we show that it is severely oversized in our Monte Carlo experiments where we maintain key features of the actual data. As another example, whereas HJLP emphasize the importance of skewness in fund returns, our empirical approach takes higher-order moments into account. Compared to the test statistics proposed in HJLP, we adjust the original percentile statistics in Fama and French (2010). Our adjusted statistics are likely more robust to extreme test statistics in the cross section and hence more informative about the additional question of *how many* funds are outperforming.

Our paper is organized as follows. Section I discusses the similarities and differences between Kosowski et al. (2006) and Fama and French (2010). Section II describes our simulation framework and presents our results. Section III addresses issues related to our simulation framework. Section IV concludes.

I. Methodological Similarities and Differences

A. Similarities

Both Kosowski et al. (2006, hereafter KTWW) and Fama and French (2010, hereafter FF) address the question of whether outperforming funds exist. Note that this question is in absolute terms (i.e., a single outperformer, if detected, provides a definitive yes to the question) and thus is different from the

[2] Also see more recent discussions in MacKinnon (2009) and Horowitz (2019).

[3] Related bootstrap techniques that adjust for serial correlation and potentially cross-sectional dependence include Politis and Romano (1994), Li and Maddala (1996), Buhlmann (1997, 1998), Lahiri (1999), Politis and White (2004), Romano, Shaikh, and Wolf (2008), and Giacomini, Politis, and White (2013). Also see the review paper by MacKinnon (2002). Different from these papers, we focus on the implementations of Kosowski et al. (2006) and Fama and French (2010)—two bootstrapping techniques that are specifically used for fund performance evaluation.

next-step question of how many funds outperform, which is also extensively studied in the literature (see, e.g., Barras, Scaillet, and Wermers (2010, 2022), Ferson and Chen (2020), and Harvey and Liu (2018)). The corresponding null hypothesis is that all funds generate a zero alpha.

Driven by this common null hypothesis, both KTWW and FF construct their tests by forcing this null to hold exactly in-sample. For our replication of these papers, we subtract the estimated alpha from each fund to obtain a pseudo panel of funds that have an in-sample alpha of exactly zero. We then treat this as the return population and resample to generate the cross section of test statistics (i.e., t-statistics) under the null hypothesis. To summarize information in the cross section, we focus on extreme percentiles (e.g., the 90$^{\text{th}}$ percentile) of the cross section of test statistics. The bootstrap allows us to obtain the null (empirical) distribution of a percentile statistic. If this percentile statistic for the actual data is too large to be explained by the null distribution, we reject the null and conclude that some fund managers must possess skill. Skill in our context is measured by after-fee excess returns.

Throughout our paper, we follow KTWW's and FF's main specifications and use the Carhart (1997) four-factor model as the benchmark model to risk-adjusted fund returns.

B. Differences

There are two main differences between KTWW's implementation and FF's implementation of the bootstrap idea: sample selection and the bootstrap approach. In the two subsections below, we first illustrate the potential impact of sample selection by examining exemplar funds (Section I.B.1). We then categorize bootstrap methods used by KTWW and FF as well as two extended approaches (Section I.B.2).

B.1. Sample Selection

FF differ from KTWW in terms of the cross section of funds that they focus on. While FF examine all funds that have at least eight observations,[4] KTWW use a more stringent threshold of 60 observations in various specifications of their paper. We illustrate the potential impact of sample length in this section, leaving more detailed power analysis to subsequent sections. In addition, sample selection may interact with the bootstrap methods, which we discuss in the next section. For now, we keep our illustration simple and focus on FF's original bootstrap approach, namely, the simultaneous bootstrap of the cross section (see Section I.B.2 for a list of alternative bootstrap methods we study).

Bootstrapping is usually performed only over the sample period for which a fund has observations (for now we refer to this as the traditional approach, which is the main approach of KTWW). FF's approach differs from the traditional approach in that they resample the entire cross section at any point

[4] See our discussion in Section III.B where we require eight distinct observations.

in time and, as such, some funds have missing observations. As a result, the number of observations for a particular fund's bootstrapped sample may differ from the number of observations in the actual sample, which may lead to a difference in the distribution of t-statistics for this approach compared to the traditional method (which does not include missing observations). FF acknowledge this difference and claim it is not a serious issue for their approach. They argue that the oversampling of some funds should roughly offset the undersampling of others, leading to a cross-sectional distribution of t-statistics that has similar properties as that generated using actual fund returns.[5]

One potential issue with FF's argument is that while it is true that the number of oversampled funds should approximately equal the number of undersampled funds in a simulation run, the impact on the individual t-statistic distributions (and hence the cross-sectional distribution of t-statistics) could be very different between oversampling and undersampling. In particular, given that a t-distribution with degrees of freedom D converges to a standard normal distribution when D is large, oversampling should not be as much of a concern as undersampling. For example, for a fund with $T = 24$ actual returns, oversampling the fund's returns (e.g., $T = 36$) is unlikely to cause a problem because both $T = 24$ and $T = 36$ generate similar distributions for the t-statistic, whereas undersampling (e.g., $T = 12$) leads to a distribution with a fatter tail than a normal distribution, which may pose a problem for the FF method.

Given FF's approach of missing data bootstrap, a low number of draws may occur for funds with short return histories. To ensure a sufficient sample size, FF require at least eight unique return observations in either the original return sample or the bootstrapped sample to include a fund in the analysis.[6] We adopt this requirement throughout our analysis.

We illustrate the asymmetric impact of oversampling and undersampling through an example. We examine the bootstrapped distribution of t-statistics for several selected funds. In particular, for a given T, we randomly select a fund with approximately T monthly observations. Focusing on this fund, we first generate the corresponding zero-alpha fund by subtracting its in-sample alpha estimate from its returns (following the FF approach) and then produce three sets of distributions by bootstrapping one million times. In the first set, we generate the distribution for the number of observations in the bootstrapped samples by following the FF approach. In the second set, we compare the bootstrapped distribution of t-statistics between the traditional approach, which we will refer to as the "complete-data" bootstrap (following KTWW), that only resamples the actual fund returns and the FF method, which we will call the "missing data" bootstrap, that resamples all time periods, including those for which the fund has missing observations. In the last set, we focus on

[5] See the third paragraph in Fama and French (2010, p. 1925).

[6] FF state that they only require eight observations, but in reality they should state that eight *unique* observations are required. We thank the referee for pointing this out. Given that many papers implemented the FF method as stated (e.g., Busse, Goyal, and Wahal (2010) and Cao et al. (2013)), in Appendix A we report the equivalent of Figure 1 with eight observations (that might not be unique). The lack of power issue is even more severe.

the FF approach by decomposing its bootstrapped distribution of t-statistics into two separate distributions, one conditional on the number of observations drawn no fewer than T (i.e., oversampling) and the other conditional on undersampling.

Figure 1 reports the results for two funds with $T \leq 24$ and Figure 2 for two funds with $24 < T \leq 60$. Let us focus on Panel B of Figure 1 first, which shows the bootstrapped distributions for a fund with $T = 23$ (i.e., roughly two years of data). The top graph (i.e., bootstrapped distribution for the number of observations) peaks at 23 and is roughly symmetric around 23. There is a large amount of variation in the bootstrapped number of observations, ranging from 11 to around 42.

The middle graph in Figure 1, Panel B (i.e., the complete-data or individual fund vs. missing-data or cross-sectional bootstrap) shows how the missing-data approach distorts the distribution of t-statistics. We focus on large realizations (i.e., t-statistics ≥ 5) of the t-statistic because they are more relevant to the FF approach, which examines the right tail of the cross-sectional distribution of t-statistics. We also winsorize the distribution at 10 to better summarize information in the right tail because the distribution of t-statistics is rather dispersed when the t-statistic is larger than 10. We observe that across all t-statistic bins, the probability generated by the FF approach (i.e., missing-data distribution) is higher than that of the complete-data distribution.

The bottom graph in Figure 1, Panel B shows the oversampling versus undersampling decomposition of the FF distribution in the middle graph. In particular, conditional on undersampling, the probability of generating a large t-statistic is uniformly larger than when we are oversampling.

Turning to Panel A, the story is more complex because the FF approach requires at least eight unique observations. This implies an asymmetric distribution (around $T = 13$, the number of observations for the original data) for the number of draws in the top graph of Panel A: the distribution is skewed to the right, implying a higher chance of oversampling than undersampling.

The middle graph of Panel A displays a similar pattern as the middle graph of Panel B: FF's missing-data bootstrap leads to a slightly higher probability of generating very large t-statistics. However, decomposing this probability into oversampling versus undersampling (as shown in the bottom graph), a different pattern emerges relative to the bottom graph of Panel B: undersampling leads to a lower chance of generating large t-statistics than oversampling. This result can be explained by the strong asymmetry in the distribution of the number of draws as required by the FF approach. Because undersampling is much less likely than oversampling, the probability of generating a given (large) t-statistic is also lower with undersampling than oversampling.[7]

[7] In Figure A1, where we require eight observations (including repeated observations), the distribution of the number of draws is less asymmetric since more draws from undersampling are acceptable (e.g., seven unique draws and one repeat would qualify for inclusion). In this case, the results for the contrast between oversampling and undersampling are similar to those in the bottom graph of Panel B: undersampling leads to a higher chance of large t-statistics across all t-statistic bins.

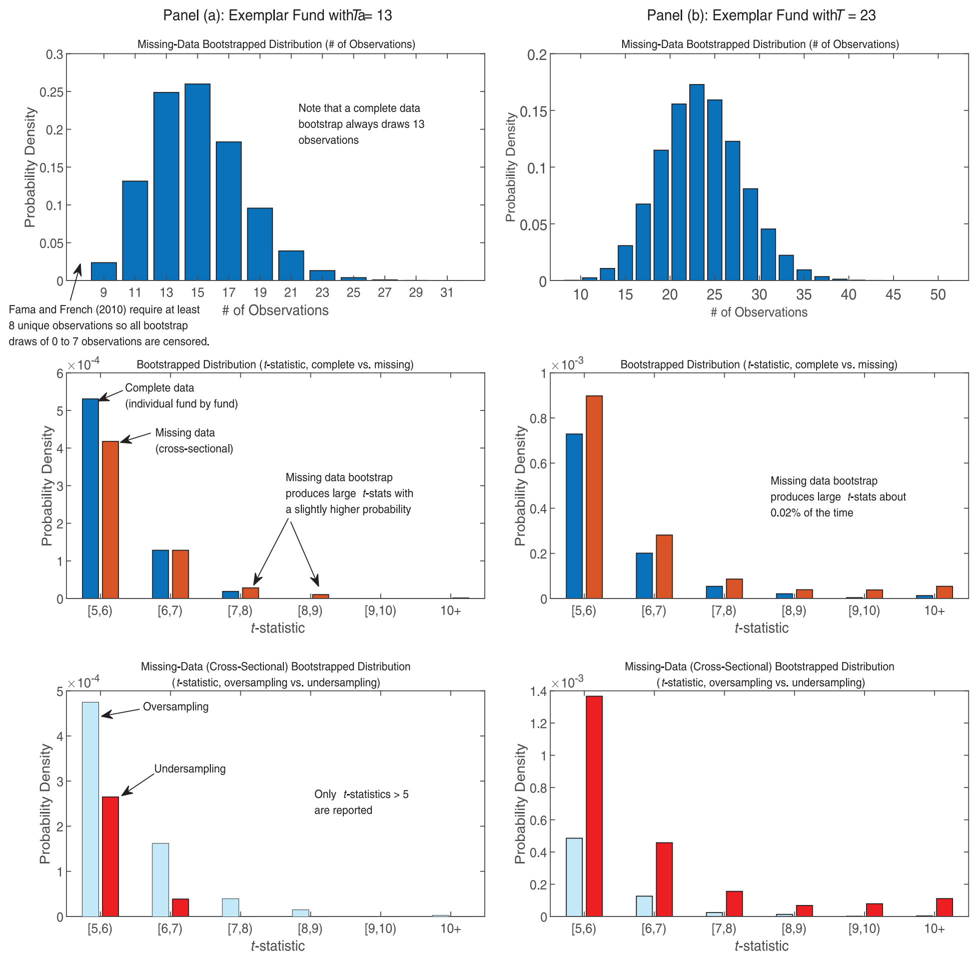

Figure 1. Bootstrapped distributions for two mutual funds with $T \leq 24$. This figure shows bootstrapped distributions for two mutual funds with $T \leq 24$. We compare the bootstrapped distributions corresponding to the "complete-data" bootstrap (individual funds) and the "missing-data" bootstrap (Fama and French (2010) or cross-sectional bootstrap). For each bootstrapping approach, we resample one million times. In each panel, we plot the bootstrapped distribution for the number of observations corresponding to the missing-data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing-data bootstrap in the bottom figure. In the top figure, the number of observations is truncated at eight based on Fama and French (2010). In the middle and bottom figures, t-statistics with a value of 5 and above are reported and truncated at 10. (Color figure can be viewed at wileyonlinelibrary.com)

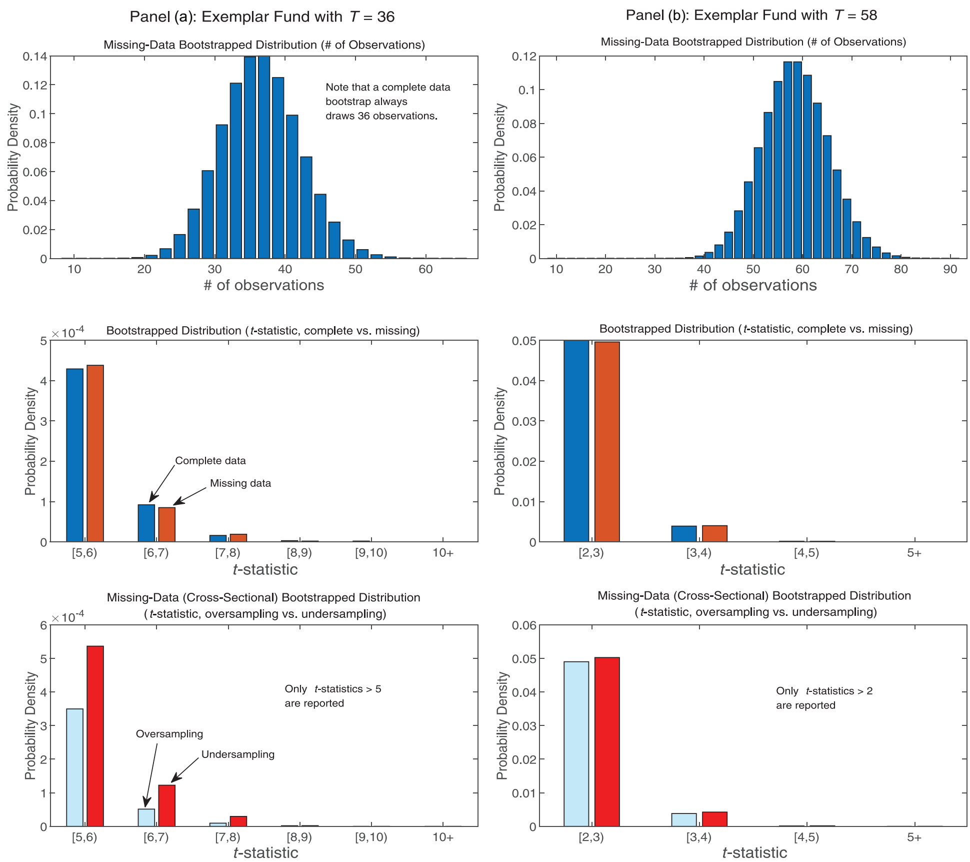

Figure 2. Bootstrapped distributions for two mutual funds with $T > 24$. This figure shows bootstrapped distributions for two mutual funds with $T > 24$. We compare the bootstrapped distributions corresponding to the "complete-data" bootstrap (individual funds) and the "missing-data" bootstrap (Fama and French (2010) or cross-sectional approach). For each bootstrapping approach, we resample one million times. In each panel, we plot the bootstrapped distribution for the number of observations corresponding to the missing-data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing-data bootstrap in the bottom figure. In the top figure, the number of observations is truncated at eight based on Fama and French (2010). In the middle and bottom figures, t-statistics with a value of 5 and above are reported and truncated at 10 in Panel A, and t-statistics with a value of 2 and above are reported and truncated at 5 in Panel B. (Color figure can be viewed at wileyonlinelibrary.com)

To summarize, we observe two patterns in Figure 1. First, FF's missing-data bootstrap has a higher chance of drawing very large t-statistics than the complete-data bootstrap. Second, everything else equal, undersampling is more likely to generate large t-statistics than oversampling. At $T = 13$, FF's approach alleviates the undersampling distortion in part by truncating the number of draws at eight (unique observations).

For larger T values, as shown in Figure 2, the difference between the complete-data bootstrap and the missing-data bootstrap is substantially smaller, although some asymmetry still exists between oversampling and undersampling for $T = 36$.

Our analysis so far focuses on the implication of the FF approach at the fund level. For a given fund, the FF bootstrapped t-statistic distribution is more fat-tailed compared to the distribution of actual returns for funds with a relatively short sample period (e.g., $T \leq 24$). This fund-level result is likely to affect FF's cross-sectional tests because the asymmetry in the bootstrapped t-statistic distribution (between oversampling and undersampling) for funds with a short history cannot be offset by funds with a larger sample, leading to a fat-tailed bootstrapped distribution for the FF test statistics (e.g., the 95$^{\text{th}}$ percentile).[8] This intuition provides the basis for our analysis of test size and power below.

To help readers navigate the statistical terms used throughout our paper, in Table I we provide a summary of the statistical terms used in the context of testing fund outperformance.

B.2. Bootstrap Methods

Figure 3 depicts the different bootstrap methods. The top panel shows the original data as well as the two individual fund "complete-data" approaches of KTWW. The bottom panels show the original cross-sectional "missing-data" approach as well as two additional approaches that mirror KTWW.

KTWW's Baseline Individual Fund Bootstrap: Residual Resampling (IND_I). KTWW's baseline bootstrap strategy resamples residuals within each fund. This is a "complete data" approach where each resampling has exactly the same number of fund observations as the historical data for the fund. In particular, for each fund, we run a factor model regression and store the regression coefficients (i.e., the alpha and factor loadings) and return residuals. At each bootstrap iteration we only sample (with replacement) individual fund residuals, which, together with the factor realizations arranged in the original chronological order and the preestimated fund betas, helps produce the pseudo-time series of fund returns. Note that α is set to zero when constructing the pseudo-time series of fund returns. We denote this bootstrap approach by IND_I where "IND" refers to individual. (See I.B.1 in KTWW for more details on this approach.) An example of this approach is presented in the middle top panel of Figure 3.

KTWW's Extended Individual Fund Bootstrap: Independent Residual and Factor Resampling (IND_{II}). To take the sampling of factors into account, KTWW also propose an extended bootstrap that features the independent

[8] Our results apply to both the left and the right tails of the cross-sectional t-statistic distribution. Given our focus on testing outperforming funds, we focus on the right tail.

Table I
Summary of Statistical Terminology

This table provides a summary of statistical terms that we use in the context of testing fund outperformance.

Terms	Description
Type I error	Assuming the null hypothesis of zero outperformance across all funds, the mistake of falsely rejecting the null and claiming outperformance for some funds.
Size	Assuming the null hypothesis of zero outperformance across all funds, the actual rate of false rejections for a given approach or the probability of making a Type I error (falsely claiming fund outperformance).
Significance level	The prespecified, desired level of size.
Type II error	Assuming the alternative hypothesis that some funds outperform is true, the mistake of not rejecting the null and falsely claiming no outperformance.
Power	Assuming the alternative hypothesis that some funds outperform is true, the actual rate of correctly identifying the existence of outperformers.
Undersampling	In a cross-sectional bootstrap (sampling a common date across all funds), undersampling occurs when the bootstrap draws fewer observations than the actual number of historical observations for the fund. This can occur because the fund does not exist in some of the months that are drawn. We refer to this as the "missing-data" bootstrap.
Oversampling	It is also possible that bootstrap could return more observations than the actual number of historical observations by oversampling months in which the fund was in existence.

resampling of factor returns and fund return residuals. This is also a complete-data approach. Similar to the baseline approach, for each fund a factor model is estimated and both regression outputs and return residuals are stored. At each bootstrap iteration, we first resample factor returns, the draws of which are the same across all funds. Then, within each fund, we resample residuals independently from the resampling of factor returns. We use both resampled residuals and resampled factor returns to construct the pseudo time series of fund returns. We denote this bootstrap approach by IND_{II}. (See Section IV.B in KTWW for more details on this approach.) Note that by keeping factor returns intact (IND_I) or resampling them simultaneously across funds (IND_{II}), the two KTWW methods preserve cross-sectional correlation in alpha caused by common factor realizations. However, they do not control for potential residual correlation as captured by FF's method.

FF's Cross-Sectional Bootstrap ($CROSS_I$). To take cross-sectional dependency into account, the FF method bootstraps time periods once at each bootstrap iteration, and the same draws of time periods apply to each fund in the cross section. Fund residuals and factor returns (which are also resampled according to the same draws of time periods) are used to construct the pseudo panel of fund returns. Think of a data matrix with time periods in rows and

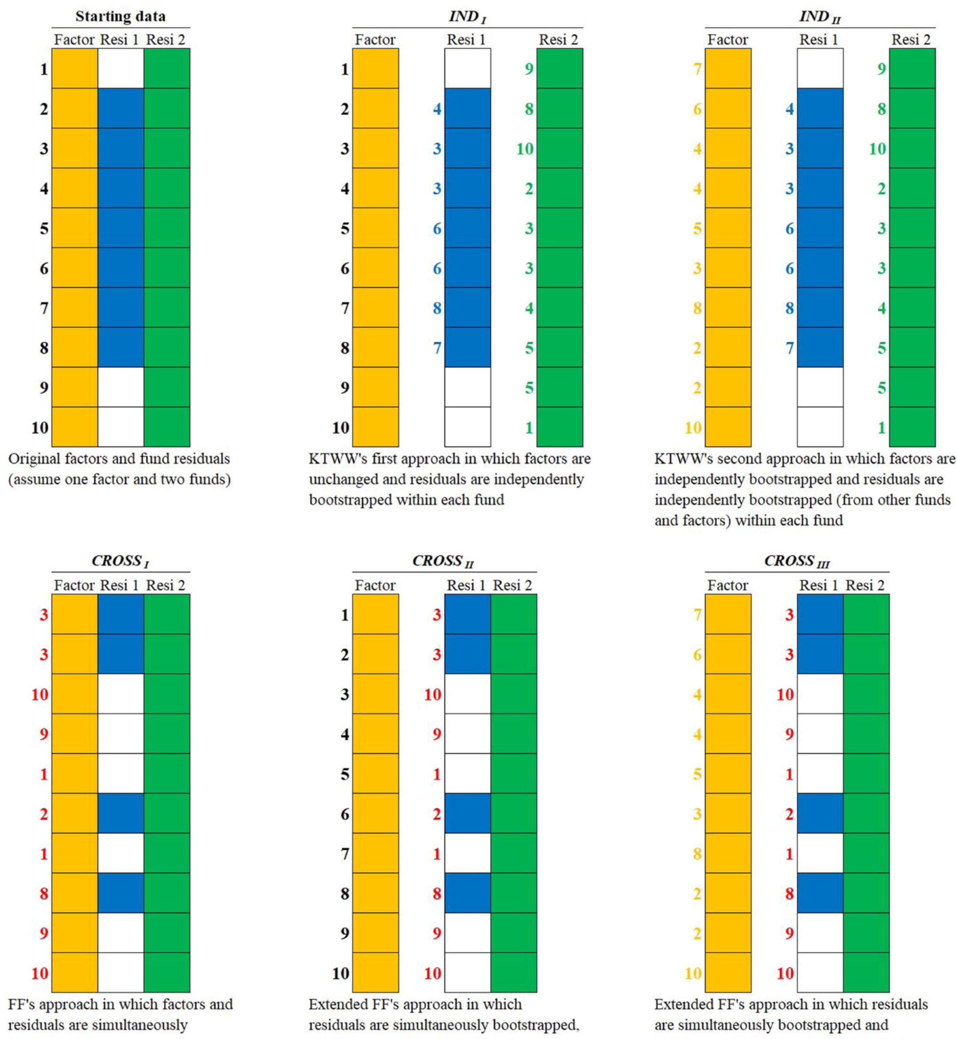

Figure 3. Five Methods: A diagrammatic display. For a fund i, let the estimated β and α be $\hat{\beta}_i$ and $\hat{\alpha}_i$, respectively. Let factor returns be F_t (assume a single factor for simplicity) and regression residuals be $\varepsilon_{i,t}$. For a bootstrap sample (after enforcing the zero-alpha assumption), we calculate the bootstrapped return according to $\hat{\beta}_i \times \tilde{F}_t + \tilde{\varepsilon}_{i,t}$, where $\tilde{F}_t$ and $\tilde{\varepsilon}_{i,t}$ are bootstrapped factor returns and residuals, respectively. Different methods amount to different ways to obtain $\tilde{F}_t$ and $\tilde{\varepsilon}_{i,t}$. We then regress bootstrapped returns on $\tilde{F}_t$ to obtain test statistics for the bootstrapped sample. (Color figure can be viewed at wileyonlinelibrary.com)

funds in columns. This method samples rows of this matrix. We denote this approach by $CROSS_I$. (See Section III.A in FF for more details on this approach.)[9]

[9] In Section IV.C, KTWW state that they implemented a similar cross-sectional approach. Given their unreported results, we mainly attribute this method to FF.

Given that time periods are drawn cross-sectionally, some observations for any given fund will be missing for funds with partial histories. This is illustrated in the bottom left panel of Figure 3.

Extended Cross-Sectional Bootstrap: CROSS_{II} *and* CROSS_{III}. The next two bootstrap approaches are not implemented by KTWW or FF, but are useful in disentangling the results of different bootstrapping methods. Both of these methods are missing-data approaches and are depicted in the last two panels of Figure 3.

The $CROSS_{II}$ approach modifies the original FF cross-sectional bootstrap approach, $CROSS_I$, by only bootstrapping fund residuals cross-sectionally at each iteration. In particular, for each fund we run a factor model regression and store the regression coefficients and return residuals. At each bootstrap iteration, we follow FF to bootstrap time periods, and the same draws of time periods apply to each fund in the cross section. We only bootstrap return residuals. These residuals, together with the factor realizations arranged in the original chronological order and the preestimated regression coefficients, generate the bootstrapped fund returns.

The $CROSS_{III}$ approach also modifies $CROSS_I$ by only bootstrapping fund residuals, but resamples factor returns independently, similar to IND_{II}. In particular, at each bootstrap iteration, we follow FF to bootstrap time periods and obtain the bootstrapped fund residuals. We then resample factor returns independently from the residual bootstrap, with the same draws of factor returns applying to each fund. Finally, we use bootstrapped fund residuals and resampled factor returns to construct the combined bootstrapped fund returns.

II. Assessing Size and Power: A Simulation Exercise

Our mutual fund data are obtained from the Center for Research in Security Prices (CRSP) Mutual Fund database after applying the same filters as in FF. The number of funds over our full sample period that have at least eight observations is 4,007. The number of funds with 24 observations or less (but at least eight observations) is 371.

A. The Simulation Design

Several challenges arise in comparing the bootstrapping methods of KTWW and FF. First, their conclusions are drawn over different samples. FF include funds with a number of observations as small as eight, whereas KTWW usually have a higher threshold for the number of fund observations. In our simulation design, we pay particular attention to this difference in sample size. Second, the FF approach is theoretically more appealing in that it controls for cross-sectional dependence in the residuals. Preserving this dependence structure in a simulation exercise is challenging if we simulate returns from a certain parametric distribution; any parametric distribution could misspecify the complex cross-sectional distribution. One novelty of our simulation design is the use of bootstrapping to overcome this issue. To be clear, although both KTWW and FF use bootstrapping, they use it to make inference. Under our simulation design, we use bootstrapping to simulate the underlying data-generating process.

Figure 4 illustrates our approach using a simplified example. In this example, there are eight funds and 15 observations. Four of the funds have complete data, two funds have 10 observations, and the final two funds have five observations. We call this the original data, $\mathcal{D}$ (see top left panel). To set up our simulation and to provide apples-to-apples comparisons with KTWW, we focus on the four funds with complete data, $\mathcal{D}^{sub}$ (see top right panel). As we discuss more below, our idea is to work with this subset of complete-data funds but intentionally drop some observations from some of the funds to recreate the distribution of history length in the original data $\mathcal{D}$. In general, $\mathcal{D}^{sub}$ is a $T \times N$ matrix, with N the number of funds and T the number of monthly periods.

Our simulation exercise is carried out as follows:

- We randomly assign alphas to funds, as depicted in the left middle panel of Figure 4. To ensure that alphas are properly scaled based on a fund's idiosyncratic risk, we obtain the risk estimates of all funds (in $\mathcal{D}^{sub}$) and randomly select a fraction of p funds to have a positive alpha. In our example in Figure 4, $p = 0.25$ so one in four funds gets an injected alpha while all other funds have alpha set to zero. For these selected funds, an information ratio, IR, is assigned to each fund, implying an alpha of $IR \times \hat{\sigma}_i$, where $\hat{\sigma}_i$ is fund i's idiosyncratic risk estimate. For the remaining funds, we set the alpha to zero so the null hypothesis holds for these funds. Let the adjusted data matrix be $\mathcal{D}_m$, where m stands for the number of iterations of random alpha assignment. The data matrix $\mathcal{D}_m$ thus contains the return population for N funds, of which $p \times N$ have an information ratio of IR and $(1 - p) \times N$ have a zero alpha.
- Note that the return population $\mathcal{D}_m$ contains $(1 - p) \times N$ of funds that have an alpha of exactly zero (by construction). For the simulated realized data (which we refer to as the realized data), which represent draws from the underlying population, this almost never happens because, while $\mathcal{D}_m$ represents the population, the realized sample is likely different from $\mathcal{D}_m$. Note that we view $\mathcal{D}_m$ as the underlying return population and hence we draw a realized return sample from it. Since $\mathcal{D}_m$ is also simulated, we refer to the corresponding sample as the simulated realized sample. We therefore first perturb the time periods (i.e., bootstrap time periods for all funds at the same time) to generate the realized data. This is displayed in the right middle panel of Figure 4. Denote the perturbed data by $\mathcal{D}^c_m$, where c stands for "complete" in that funds in $\mathcal{D}^c_m$ have a complete set of observations (e.g., 15 in Figure 4). Below we construct subsets of $\mathcal{D}^c_m$ that include funds with fewer than 15 observations. The difference between $\mathcal{D}^c_m$ and $\mathcal{D}_m$ reflects the difference between the return population ($\mathcal{D}_m$) and the realized sample ($\mathcal{D}^c_m$). Using the same bootstrap draws of time periods, we also perturb the factor returns.

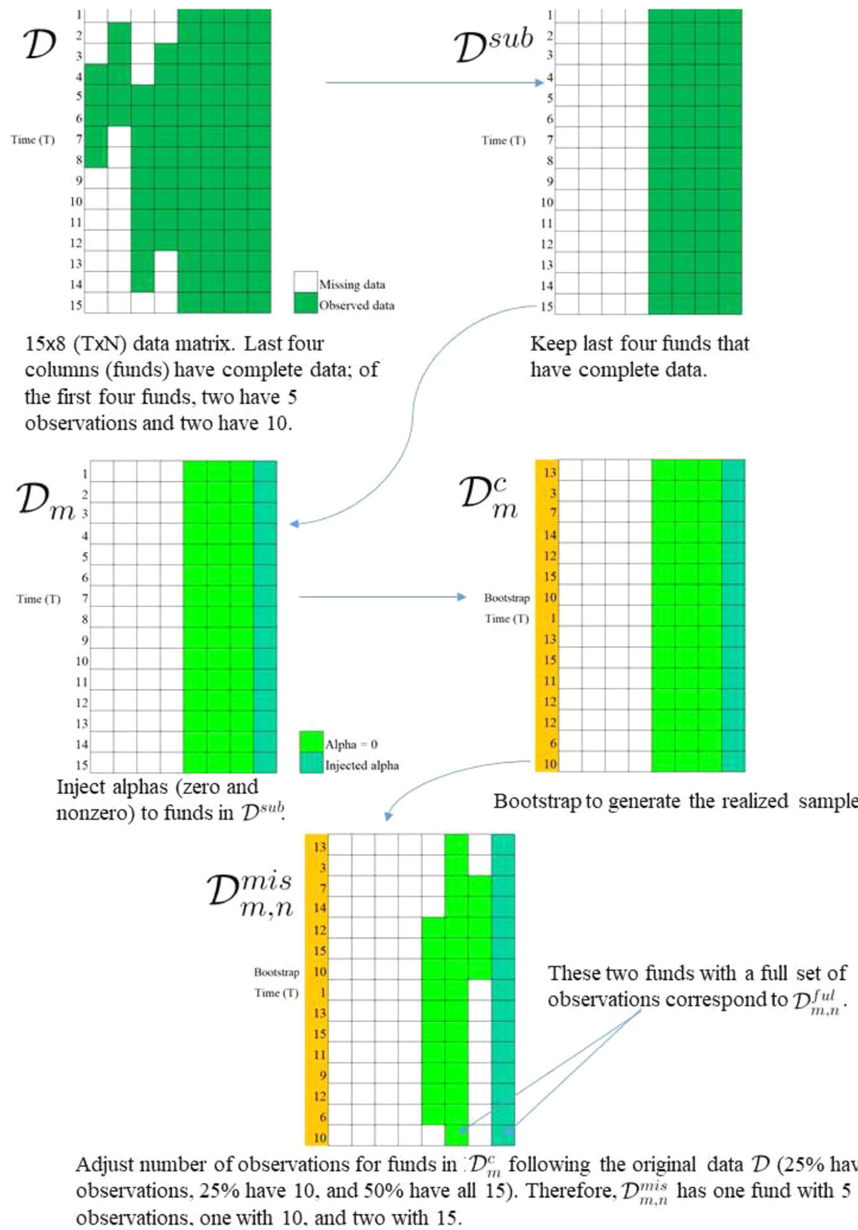

Figure 4. A visual representation of the simulation design. (Color figure can be viewed at wileyonlinelibrary.com)

We next randomly drop observations for each fund such that the empirical distribution of the cross section of the number of observations resembles the empirical distribution for the original data $\mathcal{D}$. For example, the frequency of funds with only eight observations in $\mathcal{D}$ is kept the same as in our current data. We achieve this by first obtaining the empirical distribution of the frequency of observations for the original data $\mathcal{D}$. In Figure 4 where in the original data four of eight (50%) funds have complete data, 25% of funds have one-third missing and 25% have two-thirds missing. Focusing on the four funds that have complete data, we recreate the composition of original data. In our example, two of the four funds have complete data. We delete observations for one fund so that one-third of the observations are missing and we delete two-thirds of the observations of the final fund so that our sample has the same distribution of missing values as the original data $\mathcal{D}$. Let the final data after this step be denoted by $D_{m,n}^{mis}$, where *mis* stands for missing data and n indicates the number of iterations for this step. A subsample of funds in $\mathcal{D}_{m,n}^{mis}$ has the complete history of returns (i.e., two funds in Figure 4 have $T = 15$). Let the return matrix for this subsample of funds be denoted by $\mathcal{D}_{m,n}^{ful}$, where *ful* stands for the full history of returns.

It is worth emphasizing the differences among $\mathcal{D}^{sub}$, $\mathcal{D}_m$, $\mathcal{D}_m^c$, $\mathcal{D}_{m,n}^{mis}$, and $\mathcal{D}_{m,n}^{ful}$. The data matrix $\mathcal{D}^{sub}$ includes all funds in the original data ($\mathcal{D}$) that have complete history. The data matrix $\mathcal{D}_m$ adjusts $\mathcal{D}^{sub}$ by injecting alpha into some funds and setting alpha to zero for others; it still maintains the original chronological order of time as in $\mathcal{D}$ and $\mathcal{D}^{sub}$. The data matrix $\mathcal{D}_m^c$ perturbs $\mathcal{D}_m$ (by bootstrapping the time periods) to generate the realized data. It also represents the underlying complete data that are infeasible to observe in practice, that is, it will never be the case that all funds in a particular subperiod have no missing data. The data matrix $\mathcal{D}_{m,n}^{mis}$ is a subset of $\mathcal{D}_m^c$ that includes missing observations (which we intentionally created in the data) and corresponds to the data used in FF. The data matrix $\mathcal{D}_{m,n}^{ful}$ is a subset of $\mathcal{D}_{m,n}^{mis}$ (and therefore also a subset of $\mathcal{D}_m^c$) that only contains funds with complete return history. This last data set corresponds to the sample used in KTWW.

- We have constructed three data sets ($\mathcal{D}_m^c$, $\mathcal{D}_{m,n}^{mis}$, and $\mathcal{D}_{m,n}^{ful}$) so far, and we are interested in five methods (IND_I, IND_{II}, $CROSS_I$, $CROSS_{II}$, and $CROSS_{III}$). The intersection of the two leads to 15 groups of tests. For each group, we apply the given method to one of the three data sets. Within each group, we run a host of tests that correspond to different percentiles of the cross-sectional t-statistic distribution (e.g., the maximum t-statistic and the 95$^{\text{th}}$ percentile of the t-statistics). For each test within each group, we record the testing outcome, that is, whether the null hypothesis of the nonexistence of outperforming funds is rejected for a given significance level.
- $\mathcal{D}_m^c$, $\mathcal{D}_{m,n}^{mis}$, and $\mathcal{D}_{m,n}^{ful}$ constitute the simulated panels of returns for funds. Since we know exactly which funds outperform from $\mathcal{D}_m$, we are able to

empirically evaluate the error rates for KTWW and FF. We run $M = 1,000$ (for m as in $\mathcal{D}_m$, where we randomly inject alphas into funds in $\mathcal{D}^{sub}$) and $N = 100$ (for n as in $\mathcal{D}^{mis}_{m,n}$, where we randomly drop observations from $\mathcal{D}^{c}_{m,n}$) iterations to evaluate the Type II and Type I error rates. In our context, the Type I error rate corresponds to the probability of falsely rejecting the null hypothesis. The Type II error rate corresponds to the probability of failing to reject the null hypothesis when outperforming funds exist. Test power is calculated as one minus the Type II error rate.

Similar to KTWW and FF, we run simulations for both subsamples as well as the full sample. For five-year subsamples, we examine the initial five-year period (1984 to 1988) and the last five-year period (2014 to 2018). These two periods are representative of the number of funds available in the cross section. Our simulation approach injects alphas into funds, and thus the variation in mutual fund performance over time will not affect our results. Instead, the variation in residual correlations and the number of funds available in the cross section across subsamples may have an impact.

We do not examine alternative five-year periods due to the high computational cost. The full sample covers the entire 1984 to 2018 period. Since fund sample length plays a key role in determining the performance of different bootstrapping methods, we provide a summary of fund sample length distribution in Table B.V.

B. Results for Five-Year Subsamples

B.1. Test Size

Test size is the probability of falsely rejecting the null hypothesis that all funds have zero alpha, that is, the Type I error rate (see Table I for definitions). By setting both *IR* (i.e., injected information ratio for outperforming funds) and p (i.e., assumed fraction of outperforming funds) to zero, we use our simulation framework to estimate test size. For a prespecified significance level, α, we examine how close the realized test size is in relation to α.

Figures 5 (IND_I), 6 ($CROSS_I$), and 7 ($CROSS_{II}$) include a summary of our results for the 1984 to 1988 period at the 10% significance level; Table B.I reports more detailed results. Our figures display results only for IND_I, $CROSS_I$, and $CROSS_{II}$ because of the similar performance between IND_I and IND_{II} and between $CROSS_{II}$ and $CROSS_{III}$ (Figure 3 describes the different bootstrapping methods). Our tables in Appendix B report results for all five methods.

Since we carry out our simulations under the null hypothesis, the average t-statistic and α are close to zero across the three panels in Table B.I. The maximum t-statistic shows the significance of the best-performing fund by random chance. This value is 3.06 in Panel A (funds may only have eight observations), greater than 2.67 in Panel B and 2.78 in Panel C (all funds have 60 observations in Panels B and C). These results are due to the smaller sample size for

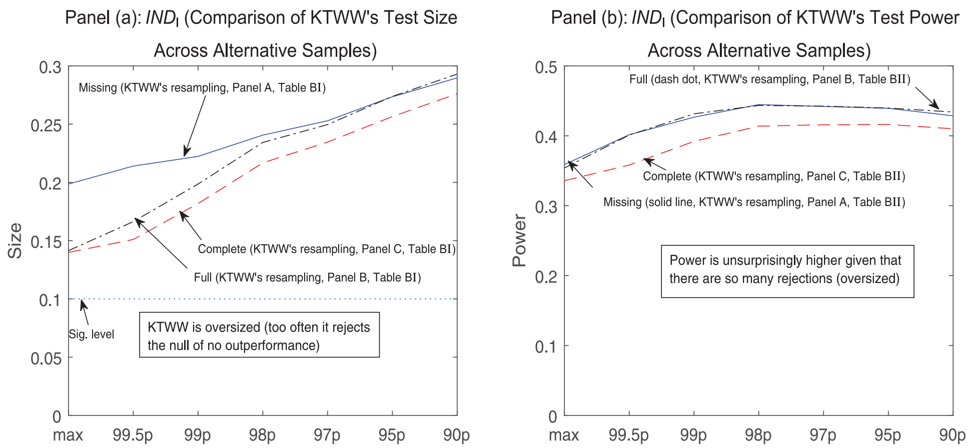

Figure 5. Results: KTWW's test size and test power, 1984 to 1988 (186 funds). We plot test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French (2010) and KTWW denotes Kosowski et al. (2006). (Color figure can be viewed at wileyonlinelibrary.com)

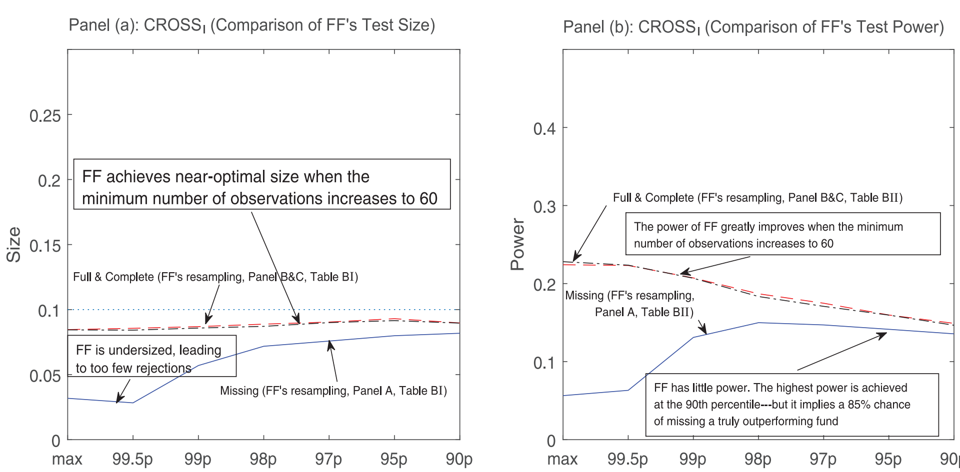

Figure 6. Results: FF's test size and test power, 1984 to 1988 (186 funds). We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French (2010) and KTWW denotes Kosowski et al. (2006). (Color figure can be viewed at wileyonlinelibrary.com)

funds in $\mathcal{D}^{mis}_{m,n}$ compared with $\mathcal{D}^{ful}_{m,n}$ and $\mathcal{D}^{c}_{m}$ (Figure 4 presents the simulation design and definitions for different data sets).

Figure 5 shows that the IND_I approach used by KTWW is substantially oversized across all three samples. All three lines (corresponding to the three samples) are well above the prespecified significance level (i.e., the dotted

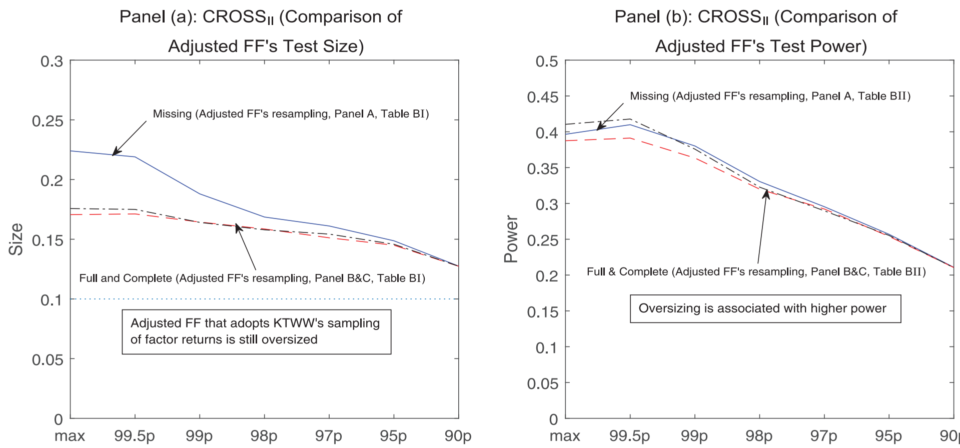

Figure 7. Results: Adjusted FF's Test Size and Test Power, 1984 to 1988 (186 Funds). We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French (2010) and KTWW denotes Kosowski et al. (2006). (Color figure can be viewed at wileyonlinelibrary.com)

benchmark line). This means that they falsely identify funds that outperform when no fund outperforms (i.e., $p = 0.0$) in the simulation. Table B.I presents detailed results for both the IND_I approach and the IND_{II} approach. For example, in Panel B of Table B.I (corresponding to KTWW's sample selection) and under 5% significance, the estimated size of KTWW's two methods (IND_I and IND_{II}) ranges from 8.5% (the max statistic) to 23.2% (the 90$^{\text{th}}$ percentile). KTWW's approaches are therefore substantially oversized for the max statistic and massively oversized for percentiles lower than, and including, the 99$^{\text{th}}$ percentile.[10]

In contrast, FF's approach ($CROSS_I$) is substantially undersized in Panel A of Table B.I, which corresponds to their application to the missing-data sample. From the perspective of hypothesis testing, undersized tests, albeit conservative (in rejecting the null), are usually regarded as acceptable because the Type I error rate constraint is satisfied. However, substantially undersized tests often lead to less powerful tests, which makes discovering outperforming funds more difficult, as we shall see below. Correspondingly, in Panel A of Figure 6, the solid line is substantially below the prespecified significance level. For instance, test size for the 99$^{\text{th}}$ percentile is only 5%. While this indicates good performance in terms of the Type I error rate, we will show that test power is low, which makes it difficult to correctly discover outperforming funds.

Figure 6 (Panel A) shows that, different from the case with missing data (i.e., Panel A in Table B.I), both the full sample, with a complete history of returns,

[10] For example, using the KTWW method, Cao et al. (2013) focus on percentiles ranging from the 90$^{\text{th}}$ to the 99$^{\text{th}}$.

and the complete sample feature size levels that are below and closer to the desired significance level.

The two modified FF approaches ($CROSS_{II}$ and $CROSS_{III}$), unlike the original FF approach $CROSS_I$, are also oversized, although usually to a lesser extent compared to the corresponding KTWW methods (see Table B.I and Figure 7).

Overall, in terms of test size, regardless of sample selection, our results suggest that nonsimultaneous sampling of factor realizations (i.e., either nonsampling of factor returns, as in IND_I and $CROSS_{II}$, or independent sampling of factor returns, as in IND_{II} and $CROSS_{III}$) leads to substantially oversized tests. This means that the null hypothesis of no outperformance is rejected too often when no fund outperforms.

B.2. Test Power

We now choose nonzero levels of p (assumed fraction of outperforming funds) and IR (injected information ratio) to study test power (see Table I for definitions). We explore nine specifications in total, with IR chosen from 0.5, 0.75, and 1.0 and p from 2.5%, 5%, and 10%. For our baseline specification, we set IR to 0.75 and p to 5%.

Figures 5, 6, and 7 also include a summary of test power and Table B.II reports more detailed results, all corresponding to our baseline specification. In Panel A of Table B.II, FF's $CROSS_I^{mis}$ approach, which corresponds to FF's sample selection, generates very low power. When 5% of outperforming funds are each endowed with an IR of 0.75, the average maximum t-statistic, α, is 3.08 (13.25%). However, the maximum power across the percentile statistics is only 15.0% at the 10% level (associated with the 98$^{\text{th}}$ percentile), implying a 85% chance of falsely claiming zero alpha across all managers.

When we alter FF's sample, as in Panels B and C, we observe a substantial increase in test power for $CROSS_I$. For example, for $\mathcal{D}_{m,n}^{ful}$ as in Panel B, test power for the maximum statistic increases to 22.4% at the 10% level. More extreme test statistics have a larger improvement in test power compared to Panel A: at the 10% level, while the power for the maximum statistic changes from 5.6% to 22.4%, the corresponding change for the 90$^{\text{th}}$ percentile is from 13.6% to 14.9%. Figure 6, Panel B, displays the difference in performance for $CROSS_I$ across samples. The two dashed lines (corresponding to $\mathcal{D}_{m,n}^{ful}$ and $\mathcal{D}_{m,n}^{c}$) dominate the solid line that corresponds to the missing sample (Figure 4 details the simulation design); these differences are smaller at lower percentiles.

The improved performance of FF's $CROSS_I$ approach, when applied to the sample with a complete history of returns (i.e., $\mathcal{D}_{m,n}^{ful}$), can be explained by the results in Table B.I. Because $CROSS_I$ is close to its optimal size when applied to $\mathcal{D}_{m,n}^{ful}$, its test power should also be high.

The two KTWW approaches (IND_I and IND_{II}) have substantially higher power than $CROSS_I$ across the three samples. However, given they are oversized, they provide ambiguous information in interpreting the test outcome because, even if the null hypothesis is rejected, it may be a false positive. The same issue applies to the two extended FF approaches ($CROSS_{II}$ and $CROSS_{III}$) (Figure 3 presents the different bootstrapping methods).

In absolute terms, test power of 22.4% (i.e., the best-case scenario for the $\mathcal{D}_{m,n}^{ful}$ sample) still seems low. This low test power says more than the general difficulty in identifying outperforming funds for the mutual fund data than about a deficiency in FF's approach. On the one hand, the close-to-optimal test size for $CROSS_{I}^{ful}$ in Table B.I is usually indicative of a powerful test. On the other hand, the large number of nonperforming funds can mask the performance of a fraction p of truly performing funds (despite an economically meaningful p and IR), leading to low power for likely any multiple testing technique that successfully guards against false positives. For instance, in Panel B of Table B.II, the average maximum t-statistic among truly performing funds is 2.94, which is not far from the average maximum t-statistic among nonperforming funds of 2.65. Moreover, the average maximum α for the former group is 11.07%, which is lower than the average maximum α for the latter group.

Tables IA.I to IA.VIII in the Internet Appendix report our results under alternative values of IR and p.[11] Not surprisingly, the highest power occurs at the maximum values of the IR and p parameters. However, even at $IR = 1$ and $p = 10\%$, the highest power is only 66.3% (at 10% level of significance) for the 99$^{\text{th}}$ percentile.

Contrary to the perception that, for a given p of the fraction of outperforming funds, the $100(1 - p)^{\text{th}}$ percentile would be most powerful (e.g., Yan and Zheng (2017)) in rejecting the null hypothesis, our results show that more extreme test statistics are usually more powerful. For instance, in the example above for Table IA.I, the highest test power is found for the 99$^{\text{th}}$ percentile, although $p = 10\%$ of funds are outperforming. In fact, test power for the 90$^{\text{th}}$ percentile is only 37.5%, substantially lower than that for the 99$^{\text{th}}$ percentile.

Overall, combining the evidence in Tables B.I and B.II, we recommend the use of the FF approach with a complete history of returns (i.e., $CROSS_{I}^{ful}$). It has near-optimal size and much higher test power compared to the case with missing observations. Among the different test statistics for $CROSS_{I}^{ful}$, we advocate the use of more extreme test statistics, such as the 99$^{\text{th}}$ percentile.

C. Results for the Full Sample

Finally, we examine the 1984 to 2018 sample. It has 2,876 funds in total.

We first clarify how we obtain the 2,876 funds. Note that our simulation design described in Section II.A cannot be directly applied because keeping funds that span the entire 1984 to 2018 period would leave us with very few funds. We adjust our simulation design as follows. First, motivated by our results in Section I.B.1, where $T = 60$ yields little distortion in the bootstrapped t-statistic distribution, we keep funds with at least 60 observations over the 1984 to 2018 period. This leaves us with 2,876 funds, which constitutes our

[11] The Internet Appendix may be found in the online version of this article.

$\mathcal{D}^{sub}$ for the 1984 to 2018 sample. Let the original sample of funds with at least eight observations be $\mathcal{D}$, which has 4,007 funds.

Second, we follow the same procedure as described in Section II.A to inject alphas into funds in $\mathcal{D}^{sub}$ and obtain the corresponding $\mathcal{D}_m$ (see Figure 4).[12] We perturb $\mathcal{D}_m$ to obtain $\mathcal{D}_m^c$. Now the question is how to insert missing observations into $\mathcal{D}_m^c$, so that the resulting data (i.e., $\mathcal{D}_m^{mis}$) have the same distribution in terms of the frequency of number of observations as $\mathcal{D}$ (i.e., the original data with 4,007 funds). We achieve this stochastically, following the idea that funds in $\mathcal{D}_m^c$ with a larger number of observations will have a higher chance of keeping more observations than funds with a lower number of observations. We calibrate our model to ensure that the frequency distribution for the number of observations for $\mathcal{D}_m^{mis}$ is approximately the same as that for $\mathcal{D}$.[13] After obtaining $\mathcal{D}_m^{mis}$, we define $\mathcal{D}_m^{ful}$ as the subsample of funds in $\mathcal{D}_m^c$ that have at least 60 observations.

Our results for the full sample are reported in Tables B.III and B.IV. Figure 8 contains a summary for the FF approach. Figures B.1 and B.2 in Appendix B contain summaries for IND_I and $CROSS_{II}$.

Figures B.1 and B.2 show that the issue of an oversized test is exacerbated for IND_I and $CROSS_{II}$ compared to the five-year subsamples. For example, as in Panel A of Figure B.1, various percentiles for IND_I reach a size around 40% when the nominal size is only 10%. In contrast, FF's $CROSS_I$ still performs well (as shown in Panel A of Figure 8): starting from the 99.5$^{\text{th}}$ percentile, although a bit oversized, all test statistics have a size close to the desired significance level. In terms of test power (Panel B of Figure 8), the preferred test statistics, such as the 99.5$^{\text{th}}$ and 99$^{\text{th}}$ percentiles, have similar but lower test power compared to the five-year subsamples (e.g., the 1984 to 1988 subsample in Figure 6). The maximum statistic is somewhat undersized and therefore less powerful than alternative test statistics.

[12] One difference from the previous five-year setting is that we need to inject a different information ratio (IR). The reason is that with the same IR, t-statistics grow in proportion to $\sqrt{T}$, where T is the number of time periods. Since our full sample has 35 years, which is seven times that over a five-year subsample, we divide the assumed IR for five-year subsamples by $\sqrt{7}$ to allow for an apples-to-apples comparison between our full-sample and subsample results. Our summary statistics reported in Table B.IV correspond well to those reported in Table B.II and Table IB.II in the Internet Appendix.

[13] For a fund with n_i observations in $\mathcal{D}_m^c$, we first record its number of observations as n_i, if $n_i < 60$. Otherwise, we randomly generate a number (denoted by p_i) from the uniform distribution between zero and one. If $p_i < a/(a + \exp(b \times (n_i - 60)))$, where a and b are our model parameters, we sample a number from $\hat{F}_{60,D}$ (i.e., the frequency distribution for the number of observations for funds in $\mathcal{D}$, conditional on funds having fewer than 60 observations) and use it as the number of observations for fund i. If $p_i \geq a/(a + \exp(b \times (n_i - 60)))$, we record the number of observations as n_i. We set the parameters a and b at 0.7 and 1/200. For $\mathcal{D}$, the mean number of observations, the probability of having fewer than 60 observations, and the standard deviation of the number of observations are 134.13, 0.28, and 97.96, respectively. The corresponding averages across simulations for $\mathcal{D}_m^{mis}$ are 139.83, 0.29, and 102.94, respectively.

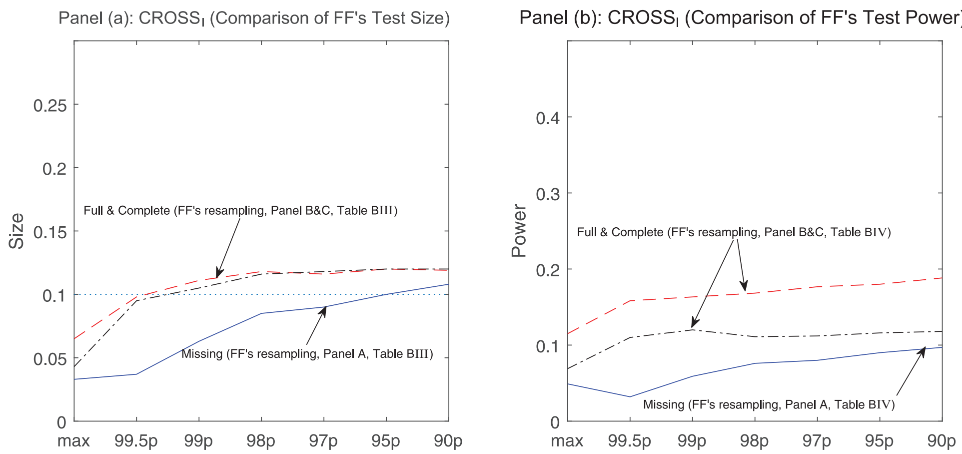

Figure 8. Results: FF's test size and test power, full sample, 1984 to 2018, 2,876 funds. We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. (Color figure can be viewed at wileyonlinelibrary.com)

D. Modifying FF: A Thresholding-FF Approach

While our strategy of keeping only those funds with more than 60 observations helps mitigate the undersampling issue of FF and enhance its test power, funds with fewer than 60 observations may represent an economically important set of funds (1,163 of 4,007 funds in our sample, which may explain FF's original intention of keeping most funds with a short return history in their paper). In particular, funds with a short history of returns may display return patterns that deviate substantially from other funds, leading to a selection bias if our goal is to make inference on the entire fund population. In this section, we propose an alternative approach that overcomes the sampling issue of FF while at the same time keeps as many funds as possible.

First, we keep all funds with a history of at least 12 monthly observations. While in principle we can keep all funds with at least eight observations, we believe 12 is a more reasonable cutoff given the increased instability of estimating t-statistics for funds with eight observations and four benchmark factors. Our thresholding-FF approach is described as follows.

Before we perform the FF bootstrap, we run a complete-sample bootstrap for each fund to generate t-statistic bandwidths that are deemed "realistic." In particular, for fund i, we subtract its in-sample alpha estimate from its returns, following FF. We then focus only on months for which we observe fund i's returns and bootstrap 1,000 times (i.e., complete-sample bootstrap). Let the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles for the bootstrapped t-statistic distribution be $\hat{q}(25, i)$ and $\hat{q}(75, i)$, respectively. The bandwidth for t-statistics that we create for fund

i is given as

$$\widehat{band}(i) = \left(\hat{q}(25,i) - thres \times [\hat{q}(75,i) - \hat{q}(25,i)], \hat{q}(75,i) + thres \times [\hat{q}(75,i) - \hat{q}(25,i)]\right),$$

where *thres* is the threshold parameter whose value is to be determined later. Note that a value of 1.5 for *thres* corresponds to the traditional rule-of-thumb for outlier detection (see, for example, Tukey (1977)). As we shall see, the optimal value of *thres* in our model is higher than 1.5, suggesting that our procedure is more conservative than the usual outlier detection rule in terms of keeping observations (i.e., more observations are classified as valid by our procedure).

Given the bandwidths for the cross section of funds, we modify the FF approach as follows. When we run FF's missing-data bootstrap (after we subtract the in-sample alphas from all funds) and for bootstrap iteration b ($b = 1, \ldots, B = 1,000$), we discard fund i if its bootstrapped t-statistic falls outside of $\widehat{band}(i)$. We discard all such funds from the cross section and compute a given percentile t-statistic (i.e., $\hat{P}_b$) based on the remaining funds. We then conduct inference by comparing the corresponding percentile for the original data with the empirical distribution $\{\hat{P}_b\}_{b=1}^{B}$.

What remains to be determined is the threshold parameter *thres*. We use our simulation approach to search for the optimal *thres* of our data. In particular, we run a grid search within the set of $thres \in \{1.0, 1.5, 2.0, \ldots, 5.0\}$. For each value of *thres*, we simulate to find test size, that is, the probability for the thresholding-FF approach to incorrectly reject the null hypothesis when the null is true. We also find the average number of funds (across bootstrapped iterations) dropped due to their extreme t-statistics in the bootstrap simulations.

Figure 9 displays our results with a significance level of 10%.[14] Not surprisingly, test size is monotonically decreasing in *thres* because the higher is *thres*, the fewer extreme t-statistic observations we drop in the bootstrapped iterations, making it harder for the FF approach to reject the null of no performance. Interestingly, all percentiles (except for the maximum t-statistic) generally achieve the desired size of 10% at $thres = 2.0$. At this value of *thres*, the average number of funds dropped in each bootstrap iteration is about 15, which is economically small compared to the size of the cross section in total (i.e., 2,876).[15]

Fixing *thres* at 2.0, Figure 10 shows test size and power for different percentile statistics. Comparing Figure 10 with the corresponding panels in Figure 8 (also labeled as "Full (FF's resampling)" in Figure 10), test size is well

[14] We choose 10% to be consistent with our previous figures. Our results are consistent across significance levels.

[15] Note that the total number of funds in $\mathcal{D}$ is greater than 2,876. However, based on our simulation design, we use $\mathcal{D}^{sub}$ to simulate the data-generating process for the panel of fund returns. The data matrix $\mathcal{D}^{sub}$ includes 2,876 funds.

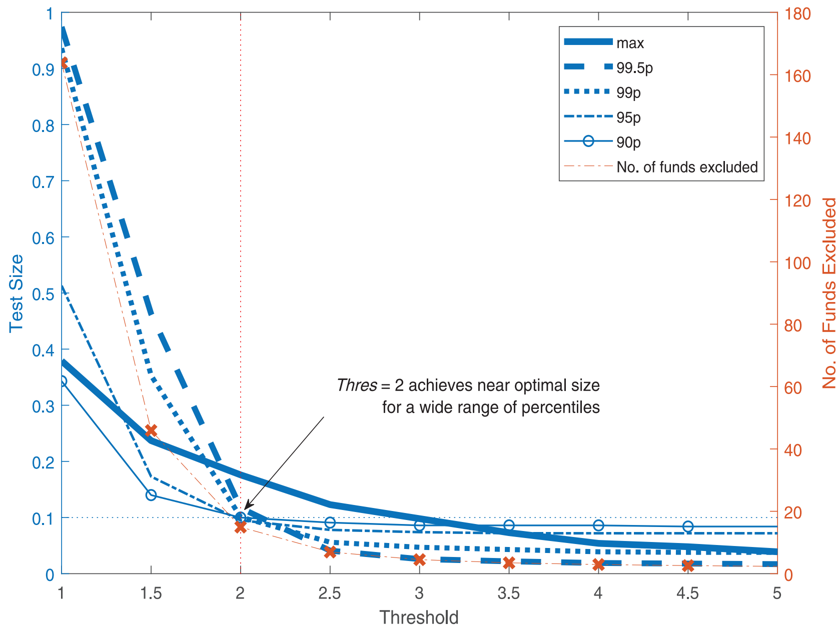

Figure 9. Results: Simulated test size for the thresholding-FF approach, 2,876 funds. We simulate to find the test size (left y-axis) for the thresholding-FF approach with a threshold parameter given by the x-axis. We also find the corresponding average number of funds dropped in the bootstrapped simulations (right y-axis). (Color figure can be viewed at wileyonlinelibrary.com)

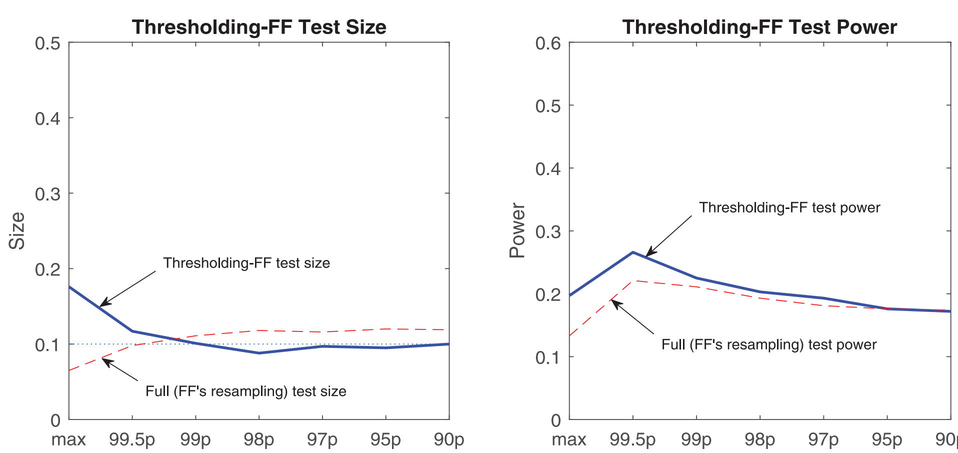

Figure 10. Results: Simulated test size and power for the thresholding-FF approach with $thres = 2.0$, **2,876 funds**. We report test size and test power at the 10% significance level for the thresholding-FF approach with the threshold parameter set to 2.0. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. (Color figure can be viewed at wileyonlinelibrary.com)

maintained at the 10% level for our thresholding-FF approach (which is also consistent with Figure 9). Meanwhile, test power is also higher, especially for more extreme percentiles such as the 99.5$^{\text{th}}$ and 99$^{\text{th}}$ percentiles. Overall, our thresholding-FF approach appears to perform well in terms of both test size and power.

Note that our results do not imply that the low-power issue for FF is caused by only 15 funds. We show that on average 15 funds are dropped across bootstrapped iterations. The total number of funds ever dropped in the bootstrapped simulations is much higher than 15. Therefore, one cannot solve the low-power issue for FF simply by excluding 15 funds from the data.

Another caveat in interpreting our results is that while $thres = 2.0$ appears to be the optimal threshold for the mutual fund data, alternative values may be found for other data sets (e.g., hedge funds) that display a different signal-to-noise ratio in the performance metric or a different dependence structure. We therefore recommend that researchers conduct similar simulation studies to find the data-specific optimal value of *thres*.

III. Other Issues

In this section, we discuss several issues related to our simulation design.

A. Alternative Five-Year Subsamples

We also examine the 2014 to 2018 sample, which features a much larger number of funds (1,502) than the 1984 to 1988 subsample (186). We report our results in Figure IB.I, Table IB.I, and Table IB.II in the Internet Appendix.

Our findings are similar to those for the 1984 to 1988 subsample. Panel C of Figure IB.I shows that overall the FF $CROSS_I$ approach performs well in test size for the full sample ($\mathcal{D}^{ful}_{m,n}$) or complete sample ($\mathcal{D}^{c}_{m,n}$). One exception is the maximum statistic, which appears to be oversized. FF do not consider the maximum statistic because it may correspond to an outlier. Our simulation reveals a similar concern: the maximum statistic in simulation runs may be too large to be explained by the bootstrapped distribution under the null, leading to overrejections. Nonetheless, starting from the 99.5$^{\text{th}}$ percentile, less extreme percentiles do not seem to be subject to this concern.

B. The Cross-Sectional Distribution of Alphas

In our simulations, we use a simple distribution to model alphas for outperforming funds. Conditional on a given p (i.e., assumed proportion of outperforming funds), we assume that all outperforming funds have the same IR. As such, we do not model the potential within-group variation in fund alphas for outperforming funds. Given the general difficulty of separating nonzero-alpha funds from zero-alpha funds, it would be even more challenging to reliably rank performance among outperforming funds. We therefore consider our simple two-group specification sufficient to approximate the cross-sectional return distribution for the underlying data-generating process.

IV. Conclusion

It is essential to attempt to separate luck from skill in the evaluation of fund performance. With so many funds, many will appear to outperform purely by luck. Bootstrapping is an attractive technique to tackle this problem and has been employed in very influential papers by Kosowski et al. (2006) and Fama and French (2010). Curiously, using similar data, they arrive at different conclusions. KTWW suggest that a measurable fraction of funds outperform while FF argue that few, if any, outperform.

Our paper replicates the findings in these papers with the goal of understanding what drives the different conclusions. We present a novel bootstrap framework that allows us to examine the Type I error rates (falsely claiming that a fund outperforms) as well as power (the probability of identifying a truly outperforming fund). In our simulation design, we know exactly which funds outperform, making it possible to measure these error rates.

There are two key differences between the KTWW and FF bootstrap implementations. First, KTWW bootstrap one fund at a time, whereas FF resample the full cross section of fund returns at every draw. Second, KTWW require a minimum of 60 observations, whereas FF require only eight time-series observations. FF's technique has the advantage of capturing economically important information in the cross section, but it also has disadvantages. Whereas the KTWW approach will always return a bootstrap simulation with the exact number of observations for the fund, the FF approach suffers from undersampling—if we start with, say, 23 fund observations, given that the cross section is being resampled, we might draw fewer than 23 observations.

Our results suggest that the undersampling of the FF approach causes problems with funds with a small number of observations. The bootstrapping technique produces very high t-statistics when there are few independent observations. These high t-statistics are inconsistent with the actual t-statistics obtained using realized data and they distort the threshold for significance. As a result, the FF implementation provides evidence that few or no funds achieve the bootstrap threshold, even when those funds have economically meaningful alphas (greater than 10% per annum). Given these results, it is perhaps unsurprising that the FF technique has little or no power to detect the truly outperforming funds in our simulation.

KTWW suffers the opposite problem. Our simulations show that KTWW substantially overrejects. This means that the KTWW approach leads researchers to falsely conclude that a large number of funds outperform.

We provide numerous simulations that are aimed at matching the particular setting that researchers face when choosing between FF and KTWW. In the end, our general recommendation is to use FF's technique that captures cross-sectional correlations, but to implement it in a way that is consistent with KTWW's approach in which the minimum number of observations is increased. For the analysis of performance, requiring a larger number of observations creates an obvious survivorship bias problem. We offer a solution using our thresholding approach. In our application, we can include funds with as few as 12 observations and achieve similar statistical performance as the approach that imposes a 60-observation minimum. Our results may alter the interpretation of published papers that use the FF or KTWW bootstrap method.

Initial submission: June 11, 2020; Accepted: January 20, 2021
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix A: Illustration: Requiring Eight Observations (Including Nonunique Observations)

We illustrate the undersampling issue that arises using the stated approach (i.e., requiring eight observations, including nonunique observations) in Fama and French (2010) and compare with our results in Figures 1 and 2 (the actual approach used in Fama and French (2010)). The stated approach is illustrated in Figures A1 and A2. This exercise is important because many researchers have implemented the stated approach. Our analysis shows that there are important differences between the stated and actual approaches for small samples.

Several patterns emerge from the comparison. First, the difference is minor for T greater than 36. The main differences stem from short-lived funds with T below 36. Second, comparing Panel A in Figure 1 and Figure A.1, the censoring implied by the actual Fama-French (2010) approach (i.e., requiring eight unique observations), as shown in Figure 1, brings the missing-data bootstrapped distribution closer to its complete-data counterpart (than Figure A.1), although the missing-data bootstrap still leads to a higher probability of very large t-statistics. The reason is that undersampling happens less frequently given the more stringent requirement on the number of unique observations. Third, comparing Panel B in Figure 1 and Figure A.1, for funds with around two years of data, it is clear that either approach tilts the missing-data bootstrapped distribution toward larger t-statistics to a greater extent than does the complete-data bootstrapped distribution. It is also evident that undersampling is driving the results.

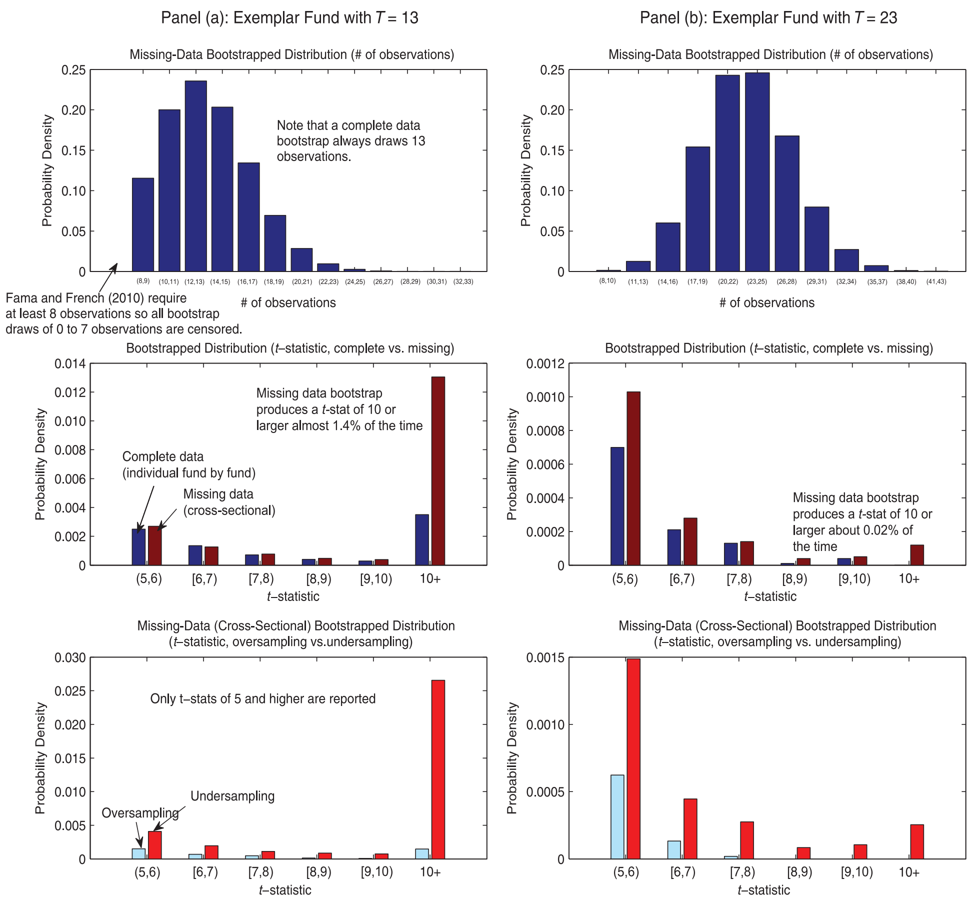

Figure A.1. Bootstrapped distributions for two mutual funds with $T \leq 24$. This figure shows bootstrapped distributions for two mutual funds with $T \leq 24$. We compare the bootstrapped distributions corresponding to the "complete-data" bootstrap (individual funds) and "missing-data" bootstrap (Fama and French (2010) or cross-sectional bootstrap). For each bootstrapping approach, we resample one million times. In each panel, we plot the bootstrapped distribution for the number of observations corresponding to the missing-data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing-data bootstrap in the bottom figure. In the top figure, the number of observations is truncated at eight based on Fama and French (2010). In the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10. We follow Fama and French's (2010) stated censoring scheme that requires eight observations (including nonunique observations). (Color figure can be viewed at wileyonlinelibrary.com)

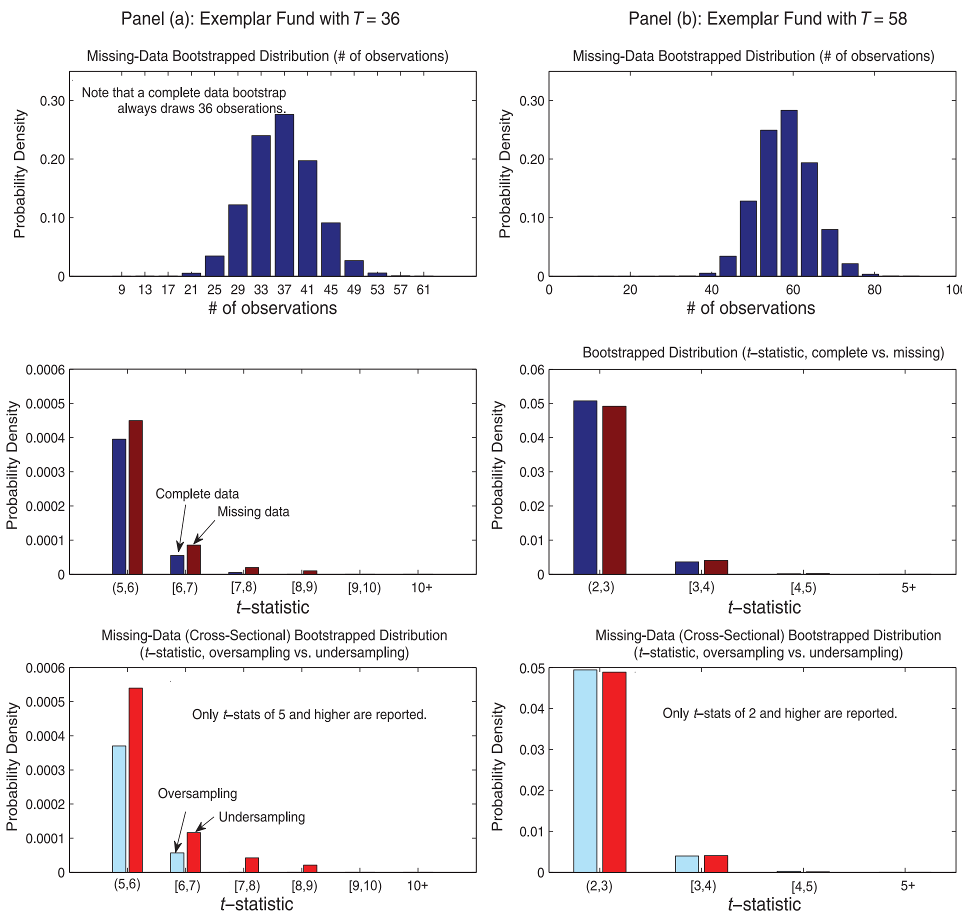

Figure A.2. Bootstrapped distributions for two mutual funds with $T > 24$. This figure shows bootstrapped distributions for two mutual funds with $T > 24$. We compare the bootstrapped distributions corresponding to the "complete-data" bootstrap (individual funds) and "missing-data" bootstrap (Fama and French (2010) or cross-sectional approach). For each bootstrapping approach, we resample one million times. In each panel, we plot the bootstrapped distribution for the number of observations corresponding to the missing-data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing-data bootstrap in the bottom figure. In the top figure, the number of observations is truncated at eight based on Fama and French (2010). In the middle and bottom figures, t-statistics with a value of 5 and above are reported and truncated at 10 for Panel A, and t-statistics with a value of 2 and above are reported and truncated at 5 for Panel B. The bin count for the top panel of Panel A for a given number c is $[c-2, c+2)$ (left close and right open). We follow Fama and French's (2010) stated censoring scheme that requires eight observations (including nonunique observations). (Color figure can be viewed at wileyonlinelibrary.com)

Appendix B: Additional Results

B.1. Five-Year Subsample, 1984 to 1988

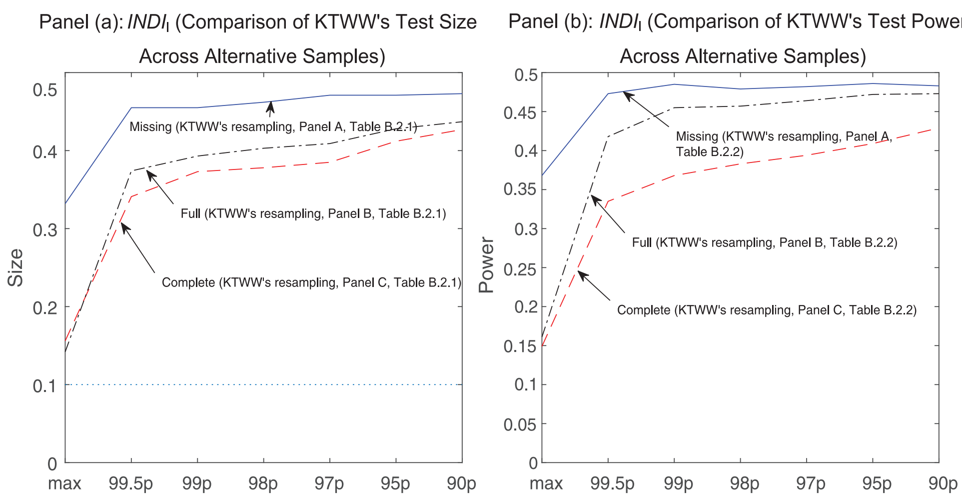

Figure B.1. Results: KTWW's test size and test power, full sample, 1984 to 2018, 2,876 funds. We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. (Color figure can be viewed at wileyonlinelibrary.com)

B.2. Full Sample, 1984 to 2018

B.3. Fund Length Distribution

Table B.I

Simulated Test Size for T = 60 (1984 to 1988)

For all funds between 1984 and 1988 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with a complete history of returns be $\mathcal{D}^{sub}$. We first inject an information ratio of $IR = 0$ into a fraction $p = 0$ of funds in $\mathcal{D}^{sub}$ and demean the remaining funds. Let the adjusted data be $\mathcal{D}_m$. For $\mathcal{D}_m$, we perturb the time periods to generate the bootstrapped sample of $\mathcal{D}^c_{m,n}$. We then randomly drop observations for funds in $\mathcal{D}^c_{m,n}$ such that the adjusted data (denoted by $\mathcal{D}^{mis}_{m,n}$) have the same cross-sectional distribution of the number of observations for each fund as $\mathcal{D}$. Let the subset of $\mathcal{D}^{mis}_{m,n}$ for which funds have a complete history of returns be given by $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$ at significance levels (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap in which return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap in which return residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap in which the same draws of time periods are used to resample both fund return residuals and factor returns at each bootstrap iteration; $CROSS_{II}$ modifies $CROSS_I$ by keeping factor returns intact (as in IND_I); and $CROSS_{III}$ modifies $CROSS_I$ by bootstrapping factor returns separately at each bootstrap iteration (as in IND_{II}). We record the testing outcome as 1 (reject) or 0 (not reject). The average rejection rate across $m = 1, 2, \ldots, 1{,}000$ and $n = 1, 2, \ldots, 100$ simulation runs generates test power. For each simulated data sample (e.g., $\mathcal{D}^{mis}_{m,n}$), we calculate summary statistics separately for funds with positive alpha (True) and zero alpha (False), such as number of funds, average (maximum) t-statistic, and average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

	Sample Statistics					Test Size							
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel A: Sample is $\mathcal{D}^{mis}_{m,n}$ (Including Missing Observations)													
# of funds	n.a.	n.a.	186.0	n.a.	IND^{mis}_{I}	*1%*	0.047	0.057	0.076	0.108	0.127	0.154	0.175
Avg. t-stat	n.a.	n.a.	−0.04	0.26		*5%*	0.129	0.146	0.158	0.185	0.205	0.228	0.249
Avg. α(%)	n.a.	n.a.	−0.07	0.87		*10%*	0.199	0.214	0.222	0.241	0.253	0.274	0.290
Max t-stat	n.a.	n.a.	3.06	0.97									
Max α (%)	n.a.	n.a.	21.76	13.03									
					IND^{mis}_{II}	*1%*	0.046	0.058	0.077	0.107	0.127	0.153	0.173
						5%	0.131	0.147	0.159	0.185	0.205	0.228	0.248
						10%	0.199	0.215	0.222	0.241	0.253	0.274	0.289
					CROSS^{mis}_{I}	*1%*	0.001	0.001	0.002	0.005	0.007	0.007	0.007
						5%	0.010	0.008	0.021	0.033	0.035	0.039	0.038
						10%	0.032	0.028	0.057	0.072	0.076	0.080	0.082

(*Continued*)

Table B.I—*Continued*

	Sample Statistics						Test Size						
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
					CROSS_{II}^{mis}	*1%*	0.049	0.047	0.032	0.028	0.026	0.022	0.016
						5%	0.141	0.137	0.112	0.098	0.090	0.082	0.067
						10%	0.224	0.219	0.188	0.169	0.161	0.149	0.128
					CROSS_{III}^{mis}	*1%*	0.049	0.047	0.033	0.028	0.025	0.022	0.015
					5%	0.143	0.139	0.111	0.099	0.089	0.082	0.067	
						10%	0.225	0.221	0.190	0.170	0.162	0.149	0.128
Panel B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only Funds with a Complete History of Returns)													
# of funds	n.a.	n.a.	139.4	5.9	IND_{I}^{ful}	*1%*	0.027	0.034	0.058	0.091	0.111	0.137	0.161
Avg. t-stat	n.a.	n.a.	−0.04	0.29		*5%*	0.085	0.095	0.127	0.165	0.186	0.210	0.232
Avg. α*(%)*	n.a.	n.a.	−0.05	0.85		*10%*	0.140	0.151	0.182	0.217	0.235	0.257	0.276
Max t-stat	n.a.	n.a.	2.67	0.67									
Max α *(%)*	n.a.	n.a.	14.01	7.84									
					IND_{II}^{ful}	*1%*	0.027	0.034	0.058	0.092	0.112	0.137	0.161
						5%	0.085	0.097	0.128	0.166	0.187	0.210	0.232
						10%	0.142	0.152	0.182	0.217	0.235	0.256	0.276
					CROSS_{I}^{ful}	*1%*	0.007	0.007	0.008	0.008	0.009	0.010	0.009
						5%	0.039	0.038	0.041	0.043	0.044	0.045	0.044
						10%	0.085	0.086	0.087	0.089	0.090	0.093	0.090
					CROSS_{II}^{ful}	*1%*	0.029	0.030	0.028	0.027	0.025	0.024	0.018
						5%	0.100	0.100	0.097	0.090	0.087	0.081	0.068
						10%	0.171	0.171	0.164	0.159	0.151	0.145	0.127
					CROSS_{III}^{ful}	*1%*	0.030	0.031	0.029	0.027	0.025	0.024	0.017
						5%	0.099	0.100	0.097	0.090	0.087	0.082	0.069
						10%	0.172	0.172	0.167	0.160	0.153	0.146	0.127

(*Continued*)

Table B.I—*Continued*

	Sample Statistics					Test Size								
	True		False				Test Statistics (of Various Percentiles)							
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel C: Sample is $\mathcal{D}_m^c$ (Infeasible)													
# of funds	n.a.	n.a.	186.0	n.a.	IND_I^c	*1%*	0.028	0.049	0.077	0.113	0.134	0.163	0.187
Avg. t-stat	n.a.	n.a.	−0.04	0.28		*5%*	0.087	0.114	0.146	0.184	0.205	0.231	0.253
Avg. α*(%)*	n.a.	n.a.	−0.05	0.83		*10%*	0.141	0.166	0.199	0.234	0.250	0.274	0.293
Max t-stat	n.a.	n.a.	2.78	0.67									
Max α *(%)*	n.a.	n.a.	15.24	8.18									
					IND_{II}^c	*1%*	0.028	0.049	0.078	0.112	0.134	0.163	0.186
						5%	0.087	0.115	0.147	0.186	0.205	0.232	0.253
						10%	0.142	0.168	0.199	0.234	0.251	0.274	0.293
					CROSS_I^c	*1%*	0.006	0.006	0.007	0.008	0.010	0.010	0.008
						5%	0.038	0.038	0.041	0.043	0.044	0.045	0.042
						10%	0.084	0.084	0.086	0.087	0.090	0.092	0.090
					CROSS_{II}^c	*1%*	0.031	0.030	0.029	0.027	0.026	0.024	0.016
						5%	0.103	0.103	0.097	0.091	0.087	0.082	0.067
						10%	0.176	0.175	0.164	0.158	0.154	0.146	0.127
					CROSS_{III}^c	*1%*	0.030	0.030	0.030	0.027	0.026	0.023	0.016
						5%	0.105	0.106	0.099	0.093	0.088	0.083	0.067
						10%	0.177	0.177	0.166	0.159	0.155	0.147	0.127

Table B.II

Simulated Test Power for T = 60 (1984 to 1988), Information Ratio IR = 0.75, and Fraction of Outperforming Funds p = 5%

For all funds between 1984 and 1988 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with a complete history of returns be $\mathcal{D}^{sub}$. We first inject an information ratio of 0.75 into $p = 5\%$ of funds in $\mathcal{D}^{sub}$ and demean the remaining funds. Let the adjusted data be $\mathcal{D}_m$. For $\mathcal{D}_m$, we perturb the time periods to generate the bootstrapped sample of $\mathcal{D}^c_{m,n}$. We then randomly drop observations for funds in $\mathcal{D}^c_{m,n}$ such that the adjusted data (denoted by $\mathcal{D}^{mis}_{m,n}$) have the same cross-sectional distribution of the number of observations for each fund as $\mathcal{D}$. Let the subset of $\mathcal{D}^{mis}_{m,n}$ for which funds have a complete history of returns be given by $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$ at significance levels (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap in which return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap in which return residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap in which the same draws of time periods are used to resample both fund return residuals and factor returns at each bootstrap iteration; $CROSS_{II}$ modifies $CROSS_I$ by keeping factor returns intact (as in IND_I); and $CROSS_{III}$ modifies $CROSS_I$ by bootstrapping factor returns separately at each bootstrap iteration (as in IND_{II}). We record the testing outcome as 1 (reject) or 0 (not reject). The average rejection rate across $m = 1, 2, \ldots, 1{,}000$ and $n = 1, 2, \ldots, 100$ simulation runs generates test power. For each simulated data sample (e.g., $\mathcal{D}^{mis}_{m,n}$), we calculate summary statistics separately for funds with positive alpha (True) and zero alpha (False), such as number of funds, average (maximum) t-statistic, and average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

	Sample Statistics				Test Power								
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel A: Sample is $\mathcal{D}^{mis}_{m,n}$ (Including Missing Observations)													
# of funds	9.0	n.a.	177.0	n.a.	IND^{mis}_I	*1%*	0.080	0.118	0.180	0.220	0.240	0.262	0.276
Avg. t-stat	1.42	0.46	–0.05	0.26		*5%*	0.239	0.290	0.334	0.358	0.366	0.372	0.373
Avg. α(%)	4.50	2.00	–0.10	0.86		*10%*	0.359	0.402	0.427	0.445	0.442	0.439	0.429
Max t-stat	3.08	0.81	3.02	0.95									
Max α (%)	13.25	8.44	21.26	12.66									
					IND^{mis}_{II}	*1%*	0.082	0.118	0.179	0.221	0.240	0.262	0.275
						5%	0.241	0.292	0.337	0.360	0.365	0.373	0.372
						10%	0.361	0.404	0.427	0.444	0.442	0.439	0.429
					CROSS^{mis}_I	*1%*	0.001	0.001	0.007	0.013	0.014	0.013	0.013
						5%	0.015	0.017	0.057	0.068	0.071	0.070	0.067
						10%	0.056	0.063	0.131	0.150	0.147	0.141	0.136

(*Continued*)

Table B.II—*Continued*

	Sample Statistics						Test Power						
	True		False					Test Statistics (of Various Percentiles)					
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
					CROSS_{II}^{mis}	*1%*	0.086	0.096	0.086	0.064	0.054	0.044	0.030
						5%	0.261	0.274	0.251	0.206	0.182	0.154	0.119
						10%	0.397	0.410	0.380	0.331	0.295	0.257	0.211
					CROSS_{III}^{mis}	*1%*	0.086	0.097	0.087	0.065	0.055	0.044	0.030
						5%	0.262	0.277	0.252	0.207	0.182	0.153	0.120
						10%	0.399	0.413	0.384	0.332	0.296	0.258	0.210
Panel B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only Funds with a Complete History of Returns)													
# of funds	6.7	1.3	132.7	5.8	IND_{I}^{ful}	*1%*	0.091	0.112	0.154	0.202	0.220	0.240	0.257
Avg. t-stat	1.54	0.51	−0.04	0.29		*5%*	0.231	0.256	0.298	0.331	0.344	0.349	0.353
Avg. α*(%)*	4.54	2.03	−0.07	0.85		*10%*	0.336	0.358	0.392	0.414	0.416	0.416	0.410
Max t-stat	2.94	0.81	2.65	0.67									
Max α *(%)*	11.07	7.13	13.62	7.55									
					IND_{II}^{ful}	*1%*	0.093	0.115	0.158	0.203	0.221	0.241	0.257
						5%	0.235	0.259	0.300	0.332	0.344	0.350	0.353
						10%	0.338	0.361	0.394	0.415	0.416	0.417	0.410
					CROSS_{I}^{ful}	*1%*	0.026	0.026	0.023	0.020	0.019	0.019	0.016
						5%	0.117	0.115	0.105	0.093	0.089	0.083	0.075
						10%	0.224	0.223	0.207	0.187	0.175	0.160	0.149
					CROSS_{II}^{ful}	*1%*	0.099	0.099	0.082	0.066	0.056	0.045	0.032
						5%	0.262	0.264	0.237	0.201	0.176	0.150	0.120
						10%	0.388	0.391	0.363	0.320	0.292	0.254	0.211
					CROSS_{III}^{ful}	*1%*	0.101	0.101	0.084	0.067	0.057	0.046	0.033
						5%	0.264	0.266	0.238	0.202	0.177	0.151	0.121
						10%	0.390	0.394	0.362	0.321	0.292	0.255	0.212

(*Continued*)

Table B.II—*Continued*

	Sample Statistics					Test Power							
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. Level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel C: Sample is $\mathcal{D}_m^c$ (Infeasible)													
# of funds	9.0	n.a.	177.0	n.a.	IND_I^c	*1%*	0.097	0.150	0.199	0.238	0.257	0.279	0.293
Avg. t-stat	1.54	0.46	−0.04	0.28		*5%*	0.246	0.304	0.343	0.368	0.374	0.381	0.383
Avg. α*(%)*	4.53	1.79	−0.07	0.83		*10%*	0.354	0.401	0.431	0.443	0.442	0.440	0.434
Max t-stat	3.12	0.78	2.75	0.67									
Max α *(%)*	12.32	7.52	14.86	7.90									
					IND_{II}^c	*1%*	0.100	0.153	0.201	0.240	0.256	0.279	0.293
						5%	0.249	0.305	0.343	0.368	0.374	0.381	0.383
						10%	0.356	0.404	0.432	0.444	0.442	0.440	0.434
					CROSS_I^c	*1%*	0.025	0.025	0.020	0.019	0.019	0.019	0.017
						5%	0.119	0.114	0.104	0.091	0.087	0.082	0.076
						10%	0.228	0.224	0.207	0.183	0.171	0.160	0.147
					CROSS_{II}^c	*1%*	0.105	0.105	0.083	0.063	0.055	0.047	0.032
						5%	0.280	0.281	0.244	0.200	0.177	0.152	0.121
						10%	0.410	0.418	0.376	0.323	0.290	0.255	0.210
					CROSS_{III}^c	*1%*	0.107	0.106	0.084	0.064	0.054	0.047	0.032
						5%	0.282	0.283	0.246	0.203	0.178	0.152	0.119
						10%	0.414	0.419	0.377	0.323	0.291	0.256	0.211

Table B.III
Simulated Test Size for Full Sample (1984 to 2018)

For all funds between 1984 and 2018 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with at least 60 monthly observations be $\mathcal{D}^{sub}$. We first inject an information ratio of $IR = 0$ into a fraction $p = 0$ of funds in $\mathcal{D}^{sub}$ and demean the remaining funds. Let the adjusted data be $\mathcal{D}_m$. For $\mathcal{D}_m$, we perturb the time periods to generate the bootstrapped sample of $\mathcal{D}^c_{m,n}$. We then randomly drop observations for funds in $\mathcal{D}^c_{m,n}$ such that the adjusted data (denoted by $\mathcal{D}^{mis}_{m,n}$) have the same cross-sectional distribution of the number of observations for each fund as $\mathcal{D}$. Let the subset of $\mathcal{D}^{mis}_{m,n}$ for which funds have a complete history of returns be given by $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$ at significance levels (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap in which return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap in which return residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap in which the same draws of time periods are used to resample both fund return residuals and factor returns at each bootstrap iteration; $CROSS_{II}$ modifies $CROSS_I$ by keeping factor returns intact (as in IND_I); and $CROSS_{III}$ modifies $CROSS_I$ by bootstrapping factor returns separately at each bootstrap iteration (as in IND_{II}). We record the testing outcome as 1 (reject) or 0 (not reject). The average rejection rate across $m = 1, 2, \ldots, 1,000$ and $n = 1, 2, \ldots, 100$ simulation runs generates test power. For each simulated data sample (e.g., $\mathcal{D}^{mis}_{m,n}$), we calculate summary statistics separately for funds with positive alpha (True) and zero alpha (False), such as number of funds, average (maximum) t-statistic, and average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

	Sample Statistics					Test Size							
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel A: Sample is $\mathcal{D}^{mis}_{m,n}$ (Including Missing Observations)													
# of funds	n.a.	n.a.	2876.0	n.a.	IND^{mis}_I	*1%*	0.145	0.281	0.323	0.359	0.371	0.387	0.400
Avg. t-stat	n.a.	n.a.	0.01	0.20		*5%*	0.262	0.396	0.401	0.422	0.425	0.443	0.449
Avg. α*(%)*	n.a.	n.a.	0.02	0.34		*10%*	0.332	0.455	0.455	0.462	0.471	0.471	0.473
Max t-stat	n.a.	n.a.	5.83	5.77									
Max α *(%)*	n.a.	n.a.	42.79	23.76									
					IND^{mis}_{II}	*1%*	0.148	0.284	0.318	0.362	0.371	0.390	0.399
						5%	0.268	0.395	0.405	0.418	0.429	0.444	0.453
						10%	0.330	0.462	0.457	0.461	0.468	0.471	0.474

(*Continued*)

Table B.III—*Continued*

	Sample Statistics						Test Size						
	True		False					Test Statistics (of Various Percentiles)					
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
					CROSS_I^{mis}	*1%*	0.003	0.004	0.006	0.012	0.013	0.017	0.017
						5%	0.017	0.019	0.035	0.038	0.042	0.048	0.054
						10%	0.033	0.037	0.063	0.085	0.090	0.100	0.108
					CROSS_{II}^{mis}	*1%*	0.134	0.050	0.043	0.035	0.029	0.031	0.024
						5%	0.256	0.163	0.157	0.135	0.124	0.112	0.102
						10%	0.329	0.291	0.254	0.223	0.207	0.202	0.173
					CROSS_{III}^{mis}	*1%*	0.132	0.048	0.044	0.034	0.029	0.027	0.024
						5%	0.256	0.162	0.156	0.135	0.122	0.111	0.102
						10%	0.338	0.288	0.250	0.223	0.206	0.202	0.171
Panel B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only Funds with More than 60 Observations)													
# of funds	n.a.	n.a.	2002.5	36.5	IND_I^{ful}	*1%*	0.019	0.208	0.263	0.311	0.328	0.341	0.361
Avg. t-stat	n.a.	n.a.	0.02	0.24		*5%*	0.078	0.294	0.322	0.347	0.367	0.385	0.407
Avg. α*(%)*	n.a.	n.a.	0.02	0.35		*10%*	0.156	0.341	0.373	0.378	0.385	0.412	0.427
Max t-stat	n.a.	n.a.	3.77	0.60									
Max α *(%)*	n.a.	n.a.	17.48	7.71									
					IND_{II}^{ful}	*1%*	0.020	0.202	0.26	0.304	0.326	0.342	0.365
						5%	0.079	0.298	0.323	0.343	0.366	0.385	0.407
						10%	0.163	0.341	0.372	0.379	0.385	0.415	0.428
					CROSS_I^{ful}	*1%*	0.007	0.012	0.012	0.012	0.012	0.015	0.015
						5%	0.029	0.053	0.061	0.057	0.054	0.055	0.063
						10%	0.065	0.098	0.111	0.118	0.116	0.12	0.119
					CROSS_{II}^{ful}	*1%*	0.024	0.037	0.041	0.029	0.028	0.027	0.026
						5%	0.083	0.116	0.118	0.113	0.112	0.103	0.097
						10%	0.180	0.199	0.208	0.186	0.184	0.169	0.155
					CROSS_{III}^{ful}	*1%*	0.023	0.040	0.037	0.030	0.030	0.027	0.029
						5%	0.081	0.116	0.120	0.113	0.110	0.101	0.094
						10%	0.184	0.196	0.207	0.189	0.186	0.172	0.158

(*Continued*)

Table B.III—*Continued*

	Sample Statistics					Test Size							
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel C: Sample is $\mathcal{D}_m^c$ (Infeasible)													
# of funds	n.a.	n.a.	2876.0	n.a.	IND_I^c	*1%*	0.014	0.255	0.293	0.328	0.344	0.365	0.389
Avg. t-stat	n.a.	n.a.	0.02	0.23		*5%*	0.068	0.324	0.359	0.366	0.386	0.394	0.415
Avg. α*(%)*	n.a.	n.a.	0.02	0.34		*10%*	0.142	0.374	0.393	0.403	0.409	0.428	0.437
Max t-stat	n.a.	n.a.	3.95	0.60									
Max α *(%)*	n.a.	n.a.	21.11	9.01									
					IND_{II}^c	*1%*	0.015	0.259	0.295	0.329	0.345	0.366	0.386
						5%	0.066	0.323	0.361	0.371	0.385	0.394	0.413
						10%	0.141	0.376	0.393	0.401	0.413	0.429	0.438
					CROSS_I^c	*1%*	0.001	0.009	0.015	0.016	0.017	0.017	0.016
						5%	0.013	0.049	0.054	0.052	0.053	0.056	0.062
						10%	0.043	0.095	0.105	0.116	0.118	0.12	0.120
					CROSS_{II}^c	*1%*	0.015	0.038	0.034	0.027	0.027	0.024	0.023
						5%	0.070	0.122	0.119	0.115	0.110	0.102	0.098
						10%	0.166	0.218	0.215	0.199	0.19	0.186	0.163
					CROSS_{III}^c	*1%*	0.016	0.035	0.034	0.029	0.028	0.025	0.025
						5%	0.071	0.124	0.124	0.116	0.113	0.105	0.095
						10%	0.165	0.214	0.214	0.197	0.191	0.183	0.160

Table B.IV

Simulated Test Power for Full Sample (1984 to 2018), information ratio $IR = 0.75/\sqrt{7}$, and fraction of outperforming funds $p = 5\%$

For all funds between 1984 and 2018 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with at least 60 monthly observations be $\mathcal{D}^{sub}$. We first inject an information ratio of $IR = 0.75/\sqrt{7}$ into $p = 5\%$ of funds in $\mathcal{D}^{sub}$ and demean the remaining funds. Let the adjusted data be $\mathcal{D}_m$. For $\mathcal{D}_m$, we perturb the time periods to generate the bootstrapped sample of $\mathcal{D}^c_{m,n}$. We then randomly drop observations for funds in $\mathcal{D}^c_{m,n}$ such that the adjusted data (denoted by $\mathcal{D}^{mis}_{m,n}$) have the same cross-sectional distribution of the number of observations for each fund as $\mathcal{D}$. Let the subset of $\mathcal{D}^{mis}_{m,n}$ for which funds have a complete history of returns be given by $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$ at significance levels (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap in which return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap in which return residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap in which the same draws of time periods are used to resample both fund return residuals and factor returns at each bootstrap iteration; $CROSS_{II}$ modifies $CROSS_I$ by keeping factor returns intact (as in IND_I); and $CROSS_{III}$ modifies $CROSS_I$ by bootstrapping factor returns separately at each bootstrap iteration (as in IND_{II}). We record the testing outcome as 1 (reject) or 0 (not reject). The average rejection rate across $m = 1, 2, \ldots, 1,000$ and $n = 1, 2, \ldots, 100$ simulation runs generates test power. For each simulated data sample (e.g., $\mathcal{D}^{mis}_{m,n}$), we calculate summary statistics separately for funds with positive alpha (True) and zero alpha (False), such as number of funds, average (maximum) t-statistic, and average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

	Sample Statistics				Test Power								
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel A: Sample is $\mathcal{D}^{mis}_{m,n}$ (including Missing Observations)													
# of funds	143.0	n.a.	2733.0	n.a.	IND^{mis}_{I}	*1%*	0.167	0.290	0.342	0.369	0.375	0.400	0.419
Avg. t-stat	0.90	0.23	0.01	0.21		*5%*	0.271	0.404	0.434	0.445	0.446	0.451	0.458
Avg. α*(%)*	1.55	0.44	0.00	0.34		*10%*	0.368	0.473	0.485	0.479	0.482	0.486	0.483
Max t-stat	4.10	1.15	6.08	6.17									
Max α *(%)*	18.55	10.38	41.32	21.83									
					IND^{mis}_{II}	*1%*	0.171	0.298	0.337	0.365	0.380	0.401	0.419
						5%	0.271	0.405	0.432	0.449	0.445	0.450	0.459
						10%	0.362	0.472	0.486	0.479	0.482	0.487	0.480
					CROSS^{mis}_{I}	*1%*	0.008	0.005	0.007	0.008	0.009	0.011	0.018
						5%	0.026	0.015	0.026	0.040	0.044	0.043	0.047
						10%	0.049	0.032	0.059	0.076	0.080	0.090	0.097

(*Continued*)

Table B.IV—*Continued*

	Sample Statistics					Test Power							
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
					CROSS_{II}^{mis}	*1%*	0.157	0.050	0.035	0.036	0.036	0.031	0.027
						5%	0.271	0.171	0.146	0.112	0.108	0.101	0.091
						10%	0.361	0.299	0.252	0.207	0.198	0.180	0.168
					CROSS_{III}^{mis}	*1%*	0.154	0.047	0.032	0.035	0.035	0.031	0.022
						5%	0.268	0.171	0.148	0.112	0.109	0.104	0.093
						10%	0.366	0.302	0.250	0.202	0.191	0.183	0.168
Panel B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only Funds with More than 60 Observations)													
# of funds	99.7	5.8	1900.9	35.2	IND_{I}^{ful}	*1%*	0.022	0.206	0.246	0.291	0.3090	0.328	0.367
Avg. t-stat	1.08	0.27	0.01	0.24		*5%*	0.086	0.283	0.314	0.350	0.364	0.386	0.408
Avg. α(%)	1.55	0.41	0.00	0.34		*10%*	0.149	0.335	0.368	0.383	0.394	0.409	0.429
Max t-stat	3.87	0.77	3.76	0.64									
Max α (%)	10.58	6.07	16.94	7.20									
					IND_{II}^{ful}	*1%*	0.026	0.211	0.245	0.287	0.310	0.332	0.367
						5%	0.082	0.287	0.314	0.351	0.367	0.385	0.411
						10%	0.150	0.331	0.365	0.381	0.396	0.407	0.428
					CROSS_{I}^{ful}	*1%*	0.013	0.020	0.020	0.022	0.023	0.027	0.032
						5%	0.048	0.082	0.080	0.088	0.092	0.087	0.085
						10%	0.115	0.158	0.163	0.168	0.177	0.180	0.188
					CROSS_{II}^{ful}	*1%*	0.026	0.033	0.030	0.031	0.029	0.027	0.024
						5%	0.089	0.109	0.110	0.096	0.093	0.092	0.085
						10%	0.175	0.201	0.174	0.167	0.159	0.156	0.151
					CROSS_{III}^{ful}	*1%*	0.028	0.034	0.029	0.030	0.028	0.027	0.025
						5%	0.094	0.111	0.111	0.094	0.095	0.093	0.083
						10%	0.176	0.203	0.175	0.166	0.158	0.157	0.151

(*Continued*)

Table B.IV—*Continued*

	Sample Statistics						Test Power						
	True		False				Test Statistics (of Various Percentiles)						
	Avg.	Std.	Avg.	Std.	Method	Sig. level	Max	*99.5%*	*99%*	*98%*	*97%*	*95%*	*90%*
Panel C: Sample is $\mathcal{D}_m^c$ (infeasible)													
# of funds	143.0	n.a.	2733.0	n.a.	IND_I^c	*1%*	0.036	0.289	0.334	0.370	0.391	0.411	0.427
Avg. t-stat	1.04	0.25	0.01	0.23		*5%*	0.092	0.361	0.404	0.424	0.438	0.455	0.460
Avg. α*(%)*	1.55	0.40	0.00	0.34		*10%*	0.161	0.418	0.455	0.457	0.464	0.472	0.473
Max t-stat	4.02	0.82	3.95	0.65									
Max α *(%)*	12.71	7.27	20.46	8.86									
					IND_{II}^c	*1%*	0.035	0.285	0.334	0.373	0.394	0.411	0.427
						5%	0.093	0.369	0.412	0.425	0.435	0.454	0.460
						10%	0.160	0.421	0.455	0.457	0.461	0.472	0.474
					CROSS_I^c	*1%*	0.007	0.012	0.013	0.012	0.014	0.019	0.021
						5%	0.034	0.055	0.053	0.060	0.060	0.061	0.061
						10%	0.069	0.110	0.120	0.111	0.112	0.116	0.118
					CROSS_{II}^c	*1%*	0.038	0.041	0.039	0.041	0.038	0.033	0.029
						5%	0.100	0.141	0.133	0.114	0.109	0.104	0.091
						10%	0.186	0.246	0.218	0.206	0.197	0.183	0.172
					CROSS_{III}^c	*1%*	0.039	0.038	0.035	0.038	0.036	0.032	0.029
						5%	0.100	0.145	0.130	0.112	0.104	0.104	0.089
						10%	0.186	0.242	0.222	0.205	0.200	0.178	0.169

Table B.V

Fund Length Distribution

We summarize fund time-series length distributions across different sample periods. $p(10)$, $p(50)$, and $p(90)$ denote the 10^{th}, 50^{th}, and 90^{th} percentile in time-series length, respectively. Top 10 (t-stat) and top 10 (alpha) focus on the top 10 ranked funds in terms of the t-statistic of alpha or alpha, respectively.

	Fund Length in Months				
Sample Period	*min*	$p(10)$	$p(50)$	$p(90)$	*max*
Panel A: 1984 to 1988 (248 Funds)					
All	10	35.3	60.0	60.0	60
Top 10 (t-stat)	47	48.0	60.0	60.0	60
Top 10 (alpha)	28	32.5	48.5	60.0	60
Panel B: 2014 to 2018 (2,235 Funds)					
All	8	25.0	60.0	60.0	60
Top 10 (t-stat)	8	18.8	60.0	60.0	60
Top 10 (alpha)	8	18.8	60.0	60.0	60
Panel C: 1984 to 2018 (4,007 Funds)					
All	8	26.0	118.0	278.8	420
Top 10 (t-stat)	8	8.5	50.0	319.0	325
Top 10 (alpha)	8	8.5	13.5	71.0	106

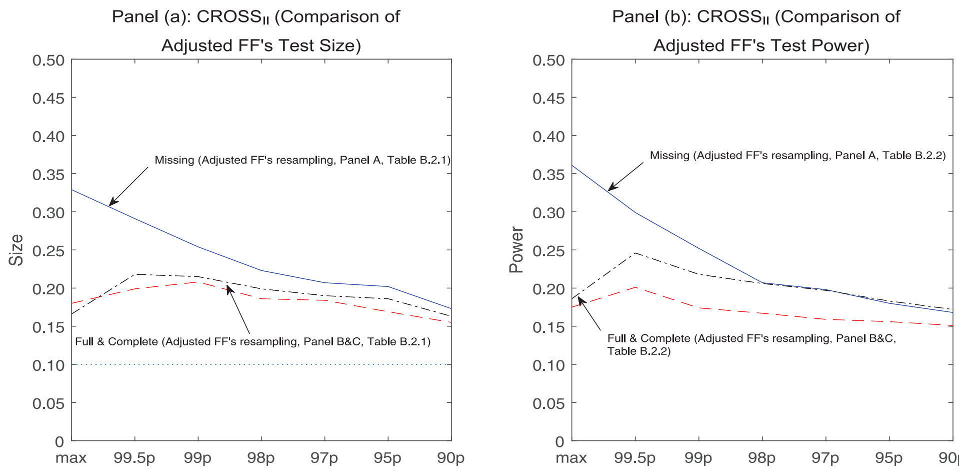

Figure B.2. Results: Adjusted FF's test size and test power, full sample, 1984 to 2018, 2,876 funds. We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. (Color figure can be viewed at wileyonlinelibrary.com)

REFERENCES

Ayadi, Mohamed A., and Lawrence Kryzanowski, 2011, Fixed-income fund performance: Role of luck and ability in tail membership, *Journal of Empirical Finance* 18, 379–392.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transaction costs, *Journal of Financial Economics* 106, 473–491.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance*, 65, 179–216.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, (2022), Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? A Reply, *Journal of Finance* 77, 601–638.

Beran, Rudolf, 1988, Prepivoting test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association* 83, 687–697.

Blake, David, Alberto Rossi, Allan Timmermann, Ian Tonks, and Russ Wermers, 2013, Decentralized investment management: Evidence from the pension fund industry, *Journal of Finance* 68, 1133–1178.

Buhlmann, Peter, 1997, Sieve bootstrap for time series, *Bernoulli* 3, 123–148.

Buhlmann, Peter, 1998, Sieve bootstrap for smoothing nonstationary time series, *Annals of Statistics* 26, 48–83.

Busse, Jeffrey A., Amit Goyal, and Sunil Wahal, 2010, Performance and persistence in institutional investment management, *Journal of Finance* 65, 765–790.

Busse, Jeffrey A., Amit Goyal, and Sunil Wahal, 2014, Investing in a global world, *Review of Finance* 18, 561–590.

Cao, Charles, Yong Chen, Bing Liang, and Andrew W. Lo, 2013, Can hedge funds time market liquidity? *Journal of Financial Economics* 109, 493–516.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chen, Yong, and Bing Liang, 2007, Do market timing hedge funds time the market? *Journal of Financial and Quantitative Analysis* 42, 827–856.

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

D'Agostino, Antonello, Kieran McQuinn, and Karl Whelan, 2012, Are some forecasters really better than others? *Journal of Money, Credit, and Banking* 44, 715–732.

Davidson, Russell, and James G. MacKinnon, 1999, The size distortion of bootstrap tests, *Econometric Theory* 15, 361–376.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance* 65, 1915–1947.

Ferson, Wayne, and Yong Chen, 2020, How many good and bad fund managers are there, really? in: Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning (World Scientific, Singapore).

Giacomini, Raffaella, Dimitris N. Politis, and Halbert White, 2013, A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators, *Econometric Theory* 3, 567–589.

Hall, Peter, 1992, The Bootstrap and Edgeworth Expansion (Springer-Verlag, New York, NY).

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

Harvey, Campbell R., and Yan Liu, 2013, Multiple testing in economics, Working paper, Duke University.

Harvey, Campbell R., and Yan Liu, 2017, Luck vs. skill and factor selection, in John Cochrane and Tobias J. Moskowitz, eds.: The Fama Portfolio (University of Chicago Press, Chicago, IL).

Harvey, Campbell R., and Yan Liu, 2018, Detecting repeatable performance, *Review of Financial Studies* 31, 2499–2552.

Harvey, Campbell R., and Yan Liu, 2020, False (and missed) discoveries in financial economics, *Journal of Finance* 75, 2503–2553.

Harvey, Campbell R., and Yan Liu, 2021, Lucky factors, *Journal of Financial Economics* 141, 413–435.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, … and the Cross-section of expected returns, *Review of Financial Studies* 29, 5–72.

Hau, Harald, and Sandy Lai, 2013, Real effects of stock underpricing, *Journal of Financial Economics* 108, 392–408.

Horowitz, Joel L., 2003, Bootstrap methods for Markov processes, *Econometrica* 71, 1049–1082.

Horowitz, Joel L., 2019, Bootstrap methods in econometrics, *Annual Review of Economics* 11, 193–224.

Huang, Haitao, Lei Jiang, Xuan Leng, and Liang Peng, 2020, Bootstrap analysis of mutual fund performance, Working paper, Georgia State University.

Jiang, George J., Tong Yao, and Tong Yu, 2007, Do mutual funds time the market? Evidence from portfolio holdings, *Journal of Financial Economics* 86, 724–758.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* 61, 2551–2595.

Lahiri, Soumendra N., 1999, Theoretical comparisons of block bootstrap methods, *Annuals of Statistics* 27, 386–404.

Li, Hongyi, and G. S. Maddala, 1996, Bootstrapping time series models, *Econometric Reviews* 15, 115–195.

MacKinnon, James G., 2002, Bootstrap inference in econometrics, *Canadian Journal of Economics* 35, 615–645.

MacKinnon, James G., 2009, Bootstrap hypothesis testing, in: Handbook of Computational Econometrics (Wiley).

Politis, Dimitris N. and Joseph P. Romano, 1994, Large sample confidence regions based on subsamples under minimal assumptions, *Journal of the American Statistical Association* 22, 2031–2050.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf, 2008, Control of the false discovery rate under dependence using the bootstrap and subsampling, *Test* 17, 417–442.

Tukey, John W., 1977, Exploratory Data Analysis (Addison-Wesley, Reading, PA).

Yan, Xuemin, and Lingling Zheng, 2017, Fundamental analysis and the cross-section of stocks returns: A data-mining approach, *Review of Financial Studies* 30, 1382–1423.

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.