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

Campbell R. Harvey
Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu*
Purdue University, West Lafayette, IN 47906 USA

Current version: October 17, 2021

Abstract

Both Kosowski et al. (2006) and Fama and French (2010) evaluate whether mutual funds outperform, but their conclusions are very different. We reconcile their findings. We show that the Fama and 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 over reject the null hypothesis, even when all funds have a zero alpha. We present a novel bootstrapping approach that should be useful to future researchers who are choosing between the two approaches.

Keywords: Performance evaluation, alpha, active management, bootstrapping, market efficiency, fund management, oversampling, undersampling, Type I errors, Type II errors

JEL:G11, G12, G23, C58

* Current Version: October 17, 2021. First posted on SSRN: June 9, 2020. We thank Ken French, Stefan Nagel (the 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.

1 Introduction

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. Two influential papers by Kosowski et al. (2006) and Fama and French (2010) employ a bootstrapping approach to try to separate luck from skill, however, they arrive at strikingly different conclusions. Kosowski et al. find that a substantial fraction of funds outperform. In contrast, Fama and French provide evidence that no advantage exists for active compared to passive management. We seek to answer the important question: Why are the conclusions 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. approach bootstraps the data firm by firm. Their approach requires a minimum of 60 observations. In contrast, Fama and French bootstrap the cross-section of fund returns and thereby retain the economically important correlation structure. Their approach requires a minimum of only 8 observations.

We provide a number of apples-to-apples comparisons of these techniques. For example, we require the Fama and French approach to have a minimum of 60 observations. We design a simulation study in which we know in advance the outperforming funds. Our technique is related to Harvey and Liu (2020) and is designed to measure the ability of each approach to correctly identify the outperforming funds. We provide five different comparisons that we believe will be useful for 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 declaring a fund to be an outperformer), Kosowski et al. is substantially over sized and therefore over rejects the market efficiency hypothesis that no fund outperforms. In contrast, the Fama and French 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 creates a strong asymmetry in the bootstrapped t-statistic distribution between undersampling and oversampling (i.e., the opposite of undersampling) for funds with a short

history, which in turn makes it difficult for their test to reject the null hypothesis. As a result, their test is under sized under the null, and consequently not powerful in detecting outperforming funds under the alternative.[1]

Reconciling these two studies, we propose two adjusted Fama and French approaches that we believe are useful for future research. The first approach simply keeps funds with a certain number of observations (e.g., 60 monthly observations) and is thus straightforward to implement. The second approach involves dropping funds with implausible t-statistics in the bootstrapped iterations. We provide guidance on what t-statistics are deemed "implausible" using our simulation approach. Both adjustments are shown to have near-optimal size and are more powerful than the original Fama and French implementation. Applying the adjusted Fama and French method, 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 present a substantial improvement over traditional approaches based on asymptotic theories for relatively small samples (see, e.g., Beran, 1998, Hall, 1992, Davidson and MacKinnon, 1999, Horowitz, 2003).[2] Despite the strong theoretical appeal, existing Monte Carlo experiments supporting bootstrap-based tests are often based on univariate tests in stylized settings. Because a generic bootstrap test that 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 carry out such an exercise for mutual fund performance evaluation, which features an unbalanced panel with a large cross section, a common factor-model benchmark across funds, and potentially a non-trivial dependence structure in fund residuals in the cross section.[3]

[1]Note 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), Chordia, Goyal, and Saretto (2020).

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

[3]Related bootstrap techniques that adjust for serial correlations 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

Another recent endeavor that aims to analyze Kosowski et al. (2006) and Fama and French (2010) is Huang, Jiang, Leng and Peng (HJLP, 2020), where they focus on the asymptotic properties of Kosowski et al. and Fama and French 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.'s approach has a correct asymptotic test size, we show 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 moment characteristics into account. Compared to the test statistics proposed in HJLP, we adjust the original percentile statistics in Fama and French, which are likely more robust to extreme test statistics in the cross section and more informative in answering the economic question: how many funds are outperforming?

Our paper is organized as follows. Section 2 discusses the commonality and differences between Kosowski et al. (2006) and Fama and French (2010). Section 3 describes our simulation framework and presents our results. Section 4 addresses several other issues related to our simulation framework. Some concluding remarks are offered in the final section.

2 Methodological Commonality and Differences

2.1 Commonality

Both Kosowski et al. (2006), hereafter KTWW, and Fama and French (2010), hereafter FF, strive to answer the same question for mutual fund performance: do outperforming funds exist? Note that the question is phrased in absolute terms (i.e., a single outperformer, if detected, provides a definitive yes to the question) and thus different from the next-step question of "how many funds outperform?" which is also extensively studied in the literature

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 particular bootstrapping techniques that are used for fund performance evaluation.

(see, e.g., Barras, Scaillet, and Wermers, 2010, 2020; 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 exactly hold 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 $90th$ 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 this null distribution, we reject the null and declare that some fund managers must possess skill. Skill in our context is measured by after fee excess returns.

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

2.2 Differences

There are two main differences between KTWW's and FF's implementation of the bootstrap idea: sample selection and the bootstrap approach. In the next two sections, we first illustrate the potential impact of sample selection by examining exemplar funds (Section 2.2.1). We then categorize the list of bootstrap methods used by KTWW and FF as well as two extended approaches (Section 2.2.2).

2.2.1 Sample Selection

FF differ from KTWW in terms of the fund cross-section they focus on. While FF keep all funds that have at least eight observations,[4] KTWW have a more stringent threshold of 60 observations in various specifications of their paper. We illustrate the potential impact of sample length in this section, while leaving more detailed power analysis to subsequent

[4]See our discussion in Section 4.2 where we require eight distinct observations.

sections. Also, 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: the simultaneous bootstrap of the cross-section (see Section 2.2.2 for a list of alternative bootstrap methods we study).

Usually, bootstrapping is performed only over the sample period for which a fund has observations (for now we term this 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 in time and, as such, some funds will have missing observations. As a result, the number of observations for the bootstrapped sample for a particular fund 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 are aware of 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, creating a cross-sectional distribution of t-statistics that has similar properties to the one generated with actual fund returns.[5]

One potential issue in 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 of 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. In contrast, 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 very low number of draws may occur for funds with short return histories. To ensure sufficient samples, FF require at least eight

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

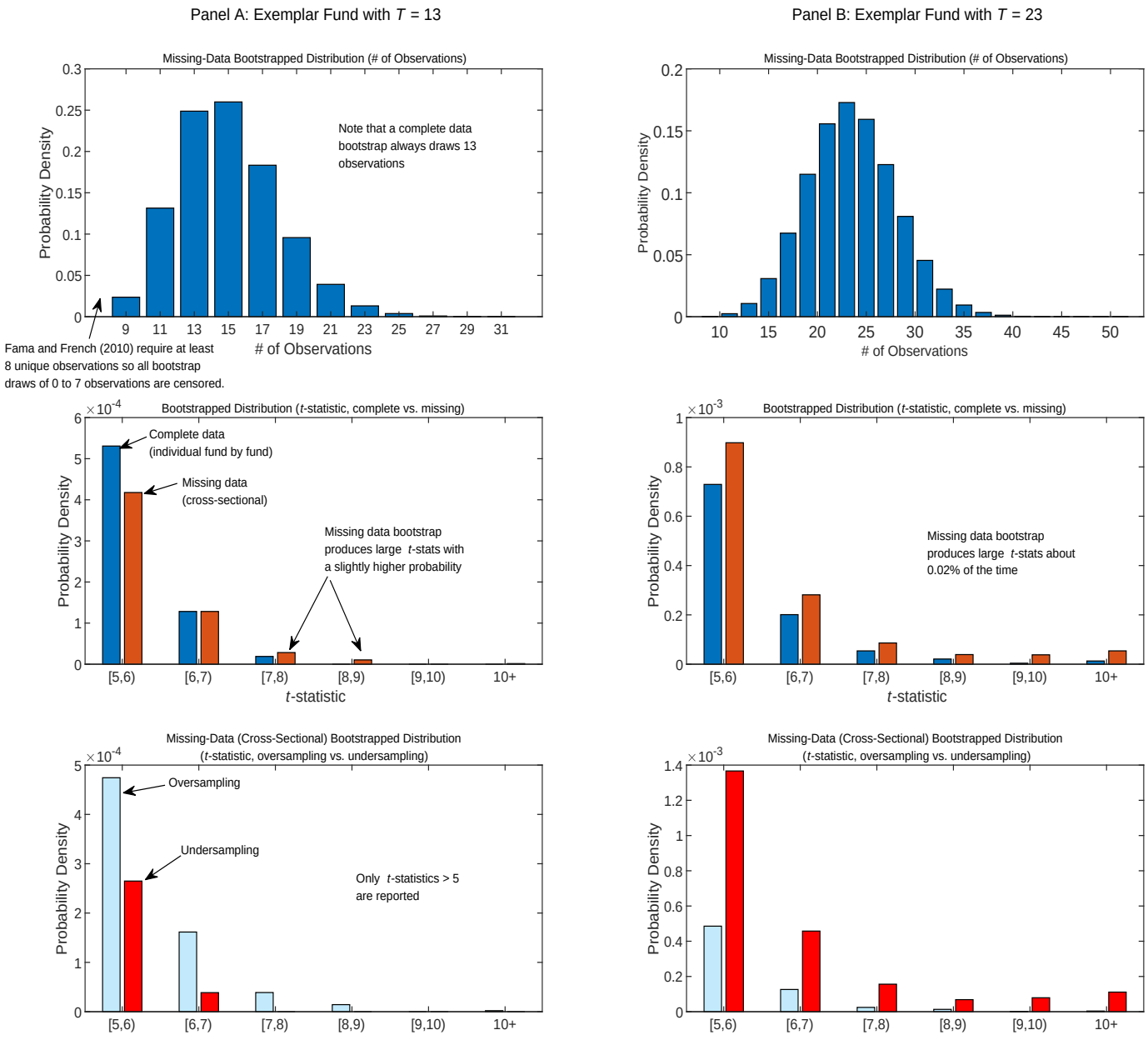
unique return observations in either the original return sample or the bootstrapped sample to include a fund in the analysis.[6] We adopt this requirement throughout our analysis.

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 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.

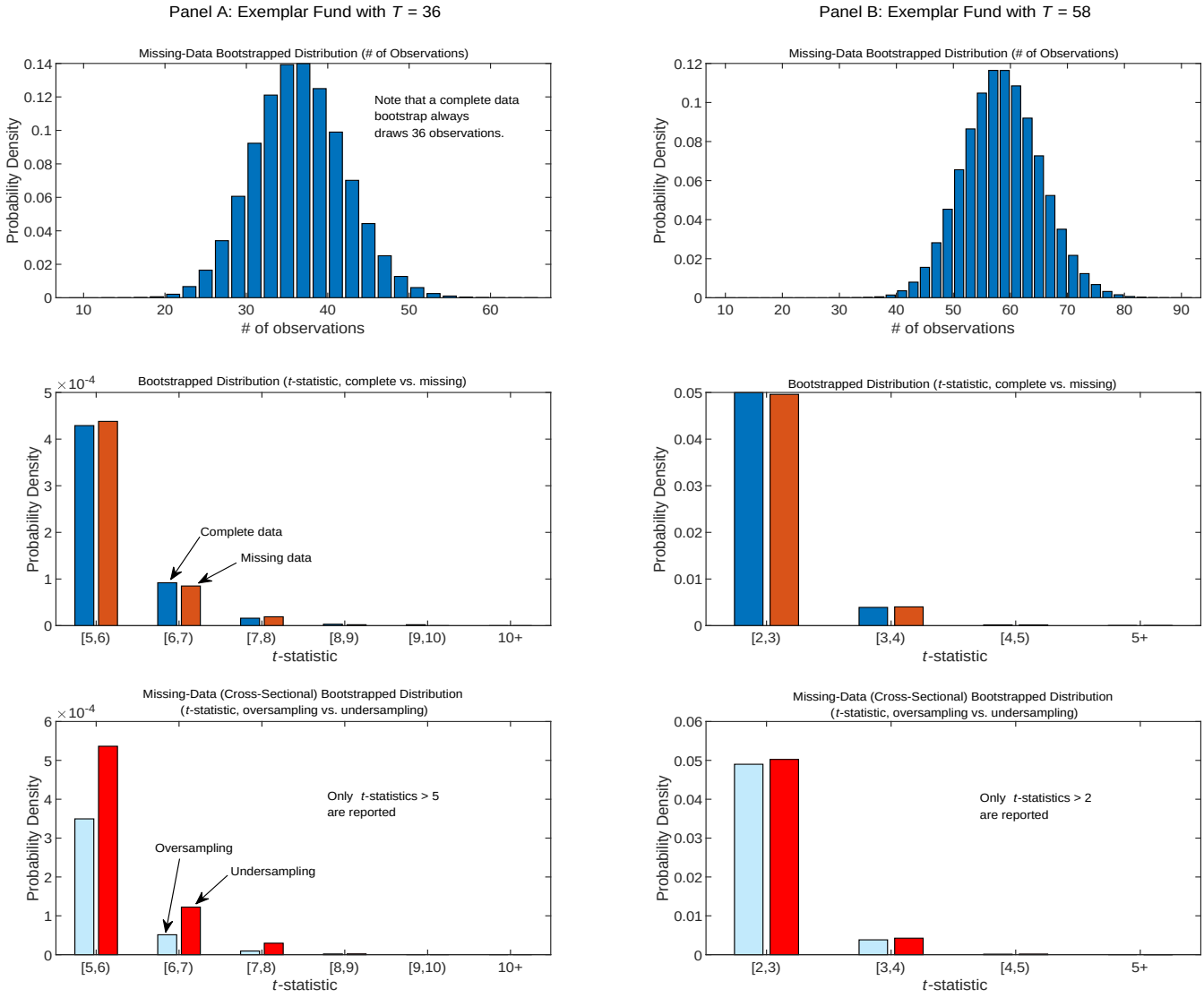
[6]Fama and French (2010) have a typo that states that they only require eight observations, whereas in reality it should be eight *unique* observations. We thank the referee for pointing this out. Given that many papers have implemented the Fama-French method as stated (e.g., Busse, Goyal, and Wahal (2010) and Cao, Chen, Liang, and Lo (2013)), we report in Appendix A the equivalent to Figure 1 with eight observations (that might not be unique). The lack of power issue is more severe.

Figure 1: **Bootstrapped Distributions for Two Mutual Funds with ($T \leq 24$)**



Bootstrapped distributions for two mutual funds with ($T \leq 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (individual fund by fund) and "missing data" bootstrap (Fama and French or cross-sectional bootstrap). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for 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. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10.

Figure 2: **Bootstrapped Distributions for Two Mutual Funds with** ($T > 24$)



Bootstrapped distributions for two mutual funds with ($T > 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (individual fund by fund) and "missing data" bootstrap (Fama and French or cross-sectional approach). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for 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. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10 for panel A, and t-statistics with a value of two and above are reported and truncated at 5 for panel B.

The middle graph in Figure 1, Panel B (i.e., the complete-data or individual fund by fund versus missing-data or cross-sectional), 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 the complete data distribution.

The bottom graph in Figure 1, Panel B, shows the oversampling versus undersampling decomposition of the FF distribution shown 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 for oversampling than undersampling.

The middle graph of Panel A displays a similar pattern to 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), we see a different pattern to the bottom graph of Panel B: undersampling leads to a lower chance of generating large t-statistics than oversampling. This can be explained by the strong asymmetry in the distribution for 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 Appendix Figure A1 where we require eight observations (including repeated observations), the distribution for 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, we show 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.

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 to some extent alleviates the undersampling distortion by truncating the number of draws at eight (unique observations).

For larger T values, as shown in Figure 2, the difference between complete data bootstrap and missing data bootstrap becomes 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 distribution of t-statistics is more fat-tailed compared to the distribution for 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^{th} percentile).[8] This intuition provides the basis for our follow-up analysis of test size and power in later sections.

To help readers better navigate the statistical terms used throughout our paper, we provide a summary of the statistical terms used in the context of testing fund outperformance in Table 1.

2.2.2 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.

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

Table 1: **Summary of Statistical Terminology**

We provide a summary of statistical terms that are used 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 pre-specified, 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	For 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 might not be in existence for some of the months that are drawn. We call this the "missing-data" bootstrap.
Oversampling	It is also possible that bootstrap could return more observations than the actual number of historical observations by oversampling particular months that the fund was in existence.

KTWW's Baseline Individual Fund Bootstrap: Residual Resampling ($\mathbf{IND}_I$) 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 pre-estimated fund betas, helps produce the pseudo-time series of fund returns. Note α is set at zero when constructing the pseudo time series of fund returns. We label this bootstrap approach IND_I where 'IND' denotes individual. (See B.1 of Section I in KTWW for more details of 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 approach that features the independent 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 the regression outputs as well as 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, residuals are resampled independently from the resampling of factor returns. Both resampled residuals and resampled factor returns are used to construct the pseudo time series of fund returns. We label this bootstrap approach IND_{II}. (See B of Section IV in KTWW for more details of this approach.) Note 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 which is 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 in rows and funds in columns. This method samples rows of this matrix. We label this approach

CROSS_I. (See A of Section III in FF for more details of this approach.)[9] 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 either 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.

CROSS_{II} 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 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 pre-estimated regression coefficients, generate the bootstrapped fund returns.

CROSS_{III} 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. Then, we resample factor returns independently from the residual bootstrap, with the same draws of factor returns applying to each fund. Lastly, bootstrapped fund residuals and resampled factor returns are used to construct the combined bootstrapped fund returns.

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

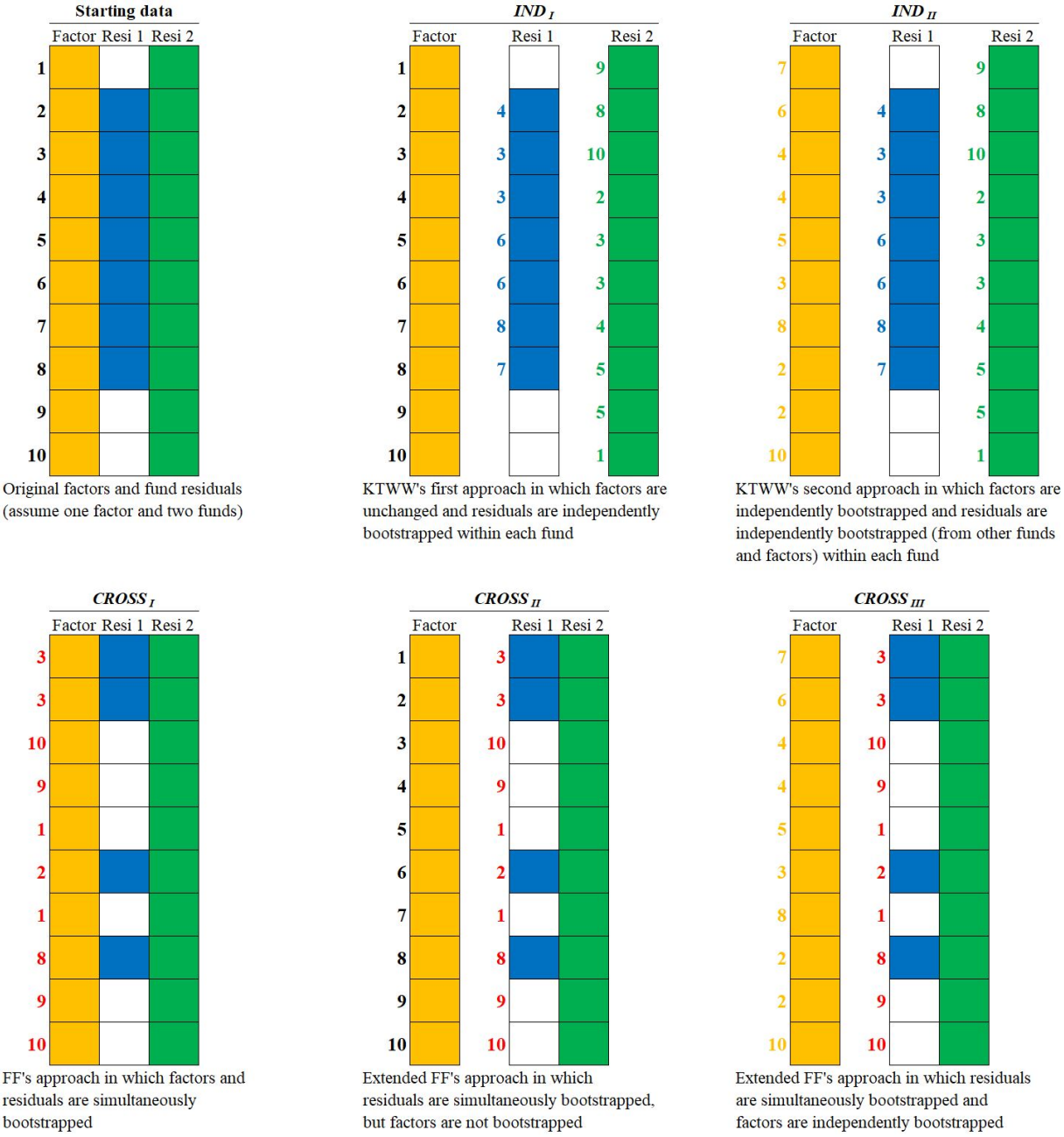


Figure 3: **Five Methods: A Diagrammatic Display.** For a fund i, let the estimated β and α be $\hat{\beta}_i$ and $\hat{\alpha}_i$, respectively. Let factor returns be F_t (assume a single factor for simplicity) and regression residuals be $\varepsilon_{i,t}$. For a bootstrap sample (after enforcing the zero-alpha assumption), we assemble the bootstrapped return as $\hat{\beta}_i \times \tilde{F}_t + \tilde{\varepsilon}_{i,t}$, 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 $\tilde{\varepsilon}_{i,t}$. Bootstrapped returns are then regressed on $\tilde{F}_t$ to obtain test statistics for the bootstrapped sample.

3 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 Fama and French (2010). 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.

3.1 The Simulation Design

Several challenges are presented by comparing the bootstrapping methods of KTWW and FF. First, their conclusions are drawn over different samples. FF include funds with the number of observations as small as eight, whereas KTWW usually have a higher threshold for the number of observations for funds. For 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 of 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 mis-specify the complex cross-sectional distribution. One novelty in 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. In our simulation design, we use bootstrapping to simulate the underlying data-generating process.

We illustrate our approach with a simplified example that is presented in Figure 4. In this example, there are 8 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, $\mathcal{D}$, and it is shown in the top left panel. To set up our simulation and to provide apples to apples comparisons with KTWW, we focus only on the four funds with complete data. We call this $\mathcal{D}^{sub}$ and it is shown in the top right panel of Figure 4. As we will describe, 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 being the number of funds and T is the number of monthly periods.

Our simulation exercise is carried out as follows:

- We randomly assign alphas to funds, which is depicted in the left middle panel of Figure 4. To make sure that alphas are properly scaled based on a fund's idiosyncratic risk, we obtain the risk estimates for all funds (in $\mathcal{D}^{sub}$) and randomly select a fraction of p funds to have a positive alpha. In our simplified example in Figure 4, $p = 0.25$ so one of 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 at 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. $\mathcal{D}_m$ thus contains the return population for N funds, among 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 a draw 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 we view $\mathcal{D}_m$ as the underlying return population and draw a realized return sample from it. Since $\mathcal{D}_m$ is also simulated, we call the corresponding sample 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. Let the perturbed data be $\mathcal{D}_m^c$, where c stands for complete in that funds in $\mathcal{D}_m^c$ have a complete set of observations (e.g., 15 in Figure 4). In contrast, we later construct subsets of $\mathcal{D}_m^c$ that include funds with fewer than 15 observations. Note the difference between $\mathcal{D}_m^c$ and $\mathcal{D}_m$ reflects the difference between the return population ($\mathcal{D}_m$) and the realized sample ($\mathcal{D}_m^c$). Using the same bootstrap draws of time periods, we also perturb the factor returns.

  We then 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 8 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%) of funds have complete data, 25% of funds have one third missing and 25% of funds have two thirds missing. Focusing only on the four funds that have complete data, we recreate the composition of original data. In our example, two of the four funds will have complete data. We delete observations for one fund so that one third of the observations are missing and 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 $D^{mis}_{m,n}$, where *mis* stands for missing data, and n indicates the number of iterations for this step. Further, a subsample of funds in $\mathcal{D}^{mis}_{m,n}$ 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 $\mathcal{D}^{ful}_{m,n}$, where *ful* stands for the full history of returns.

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

- We have constructed three data sets ($\mathcal{D}^c_m$, $\mathcal{D}^{mis}_{m,n}$, and $\mathcal{D}^{ful}_{m,n}$) 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 a 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^{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 non-existence of outperforming funds is rejected or not 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. Moreover, we know exactly which funds outperform from $\mathcal{D}_m$. As such, 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 $\mathcal{D}_{m,n}^{mis}$, where we randomly drop observations from $\mathcal{D}_{m,n}^c$) 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 and the full sample. For five-year subsamples, we examine the initial five-year (1984–1988) and the last five-year (2014–2018) period. These two periods are representative of the number of funds available in the cross-section. Our simulation approach injects alphas into funds, thus the variation in mutual fund performance across time will not have an influence on 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 1984–2018 period. Since fund sample length plays a key role in determining the performance of different bootstrapping method, we provide a summary of fund sample length distribution in Table B.3.1 in the appendix.

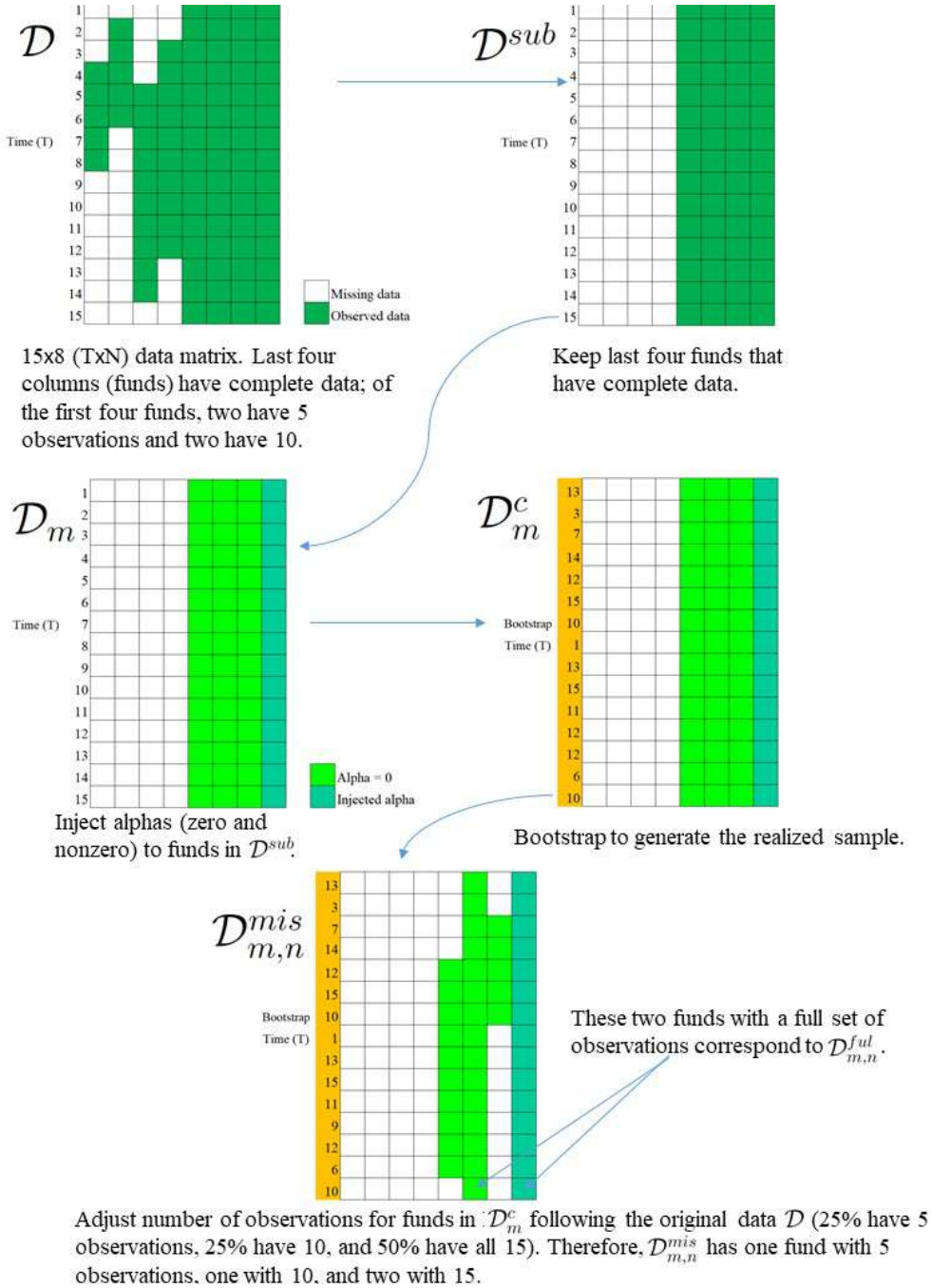


Figure 4: **A Visual Representation of the Simulation Design**

3.2 Results for Five-Year Subsamples

3.2.1 Test Size

Test size is the probability of falsely rejecting the null hypothesis that all funds have zero alpha. It is the Type I error rate (see Table 1 for definitions). By setting both IR (i.e., injected information ratio for outperforming funds) and p (i.e., the assumed fraction of outperforming funds) at zero, we use our simulation framework to estimate test size. For a pre-specified 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 1984–1988 period at the 10% significance level and Table B.1.1 in the appendix reports more detailed results. Our figures only display results 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 the appendix 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.1.1. The maximum t-statistic shows the significance of the best performing fund by random chance, which is 3.06 for Panel A (funds may only have 8 observations), higher than 2.67 for Panel B and 2.78 for Panel C (all funds have 60 observations for Panels B and C). This is caused by the smaller sample size for 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 over sized across all three samples. All three lines (corresponding to the three samples) are well above the pre-specified significance level (i.e., the dotted benchmark line). This means they falsely identify funds that outperform when no fund outperforms (i.e., $p = 0.0$) in the simulation. Table B.1.1 presents detailed results for both the IND_I and the IND_{II} approach. For example, in Panel B of Table B.1.1 (corresponding to KTWW's sample selection) and under 5% significance, the estimated size for KTWW's two methods (IND_I and IND_{II}) ranges from 8.5% (the max statistic) to 23.2% (the 90^{th} percentile). They are therefore substantially oversized for

the max statistic and massively oversized for percentiles lower than, and including, the 99^{th} percentile.[10]

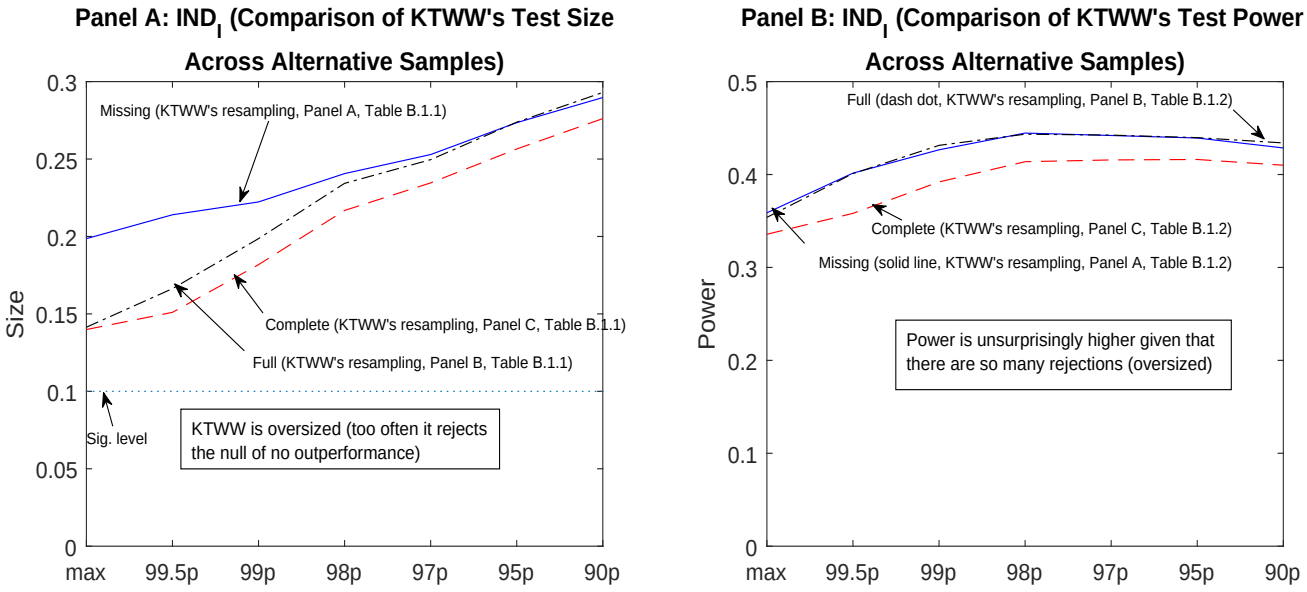


Figure 5: **Results: KTWW's Test Size and Test Power, 1984–1988 (186 Funds).** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French (2010)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

In contrast, FF's approach ($CROSS_I$) is substantially under sized in Panel A of Table B.1.1, which corresponds to their application to the missing data sample. From the perspective of hypothesis testing, under sized tests, albeit conservative (in rejecting the null), are usually regarded as acceptable because the Type I error rate constraint is satisfied. However, substantially under sized tests often lead to less powerful tests, which makes discovering outperforming funds more difficult, as we shall see later. Correspondingly, in Panel A of Figure 6, the solid line is substantially below the pre-specified significance level. For instance, test size for the 99^{th} percentile 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.

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

Figure 6 (Panel A) shows that, different from the case with missing data (i.e., Panel A in Table B.1.1), both the full sample, with a complete history of returns, and the complete sample feature size levels that are below and closer to the desired significance level.

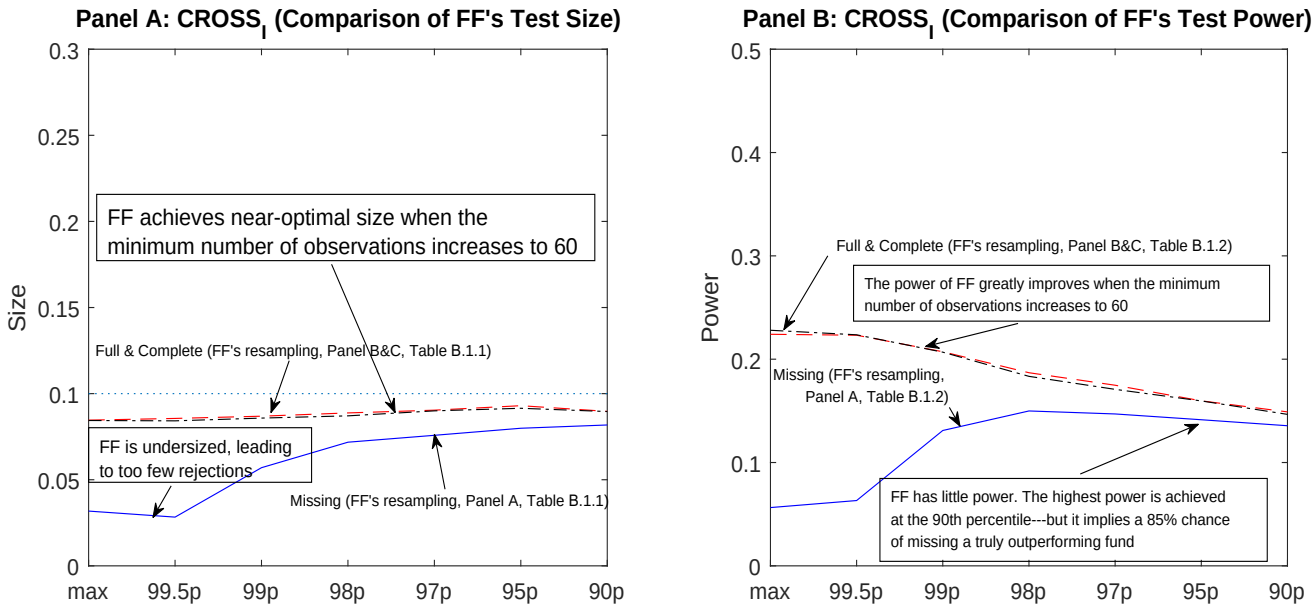


Figure 6: **Results: FF's Test Size and Test Power, 1984–1988 (186 Funds).** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French (2010) and KTWW denotes Kosowski et al. (2006).

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.1.1 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 non sampling 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.

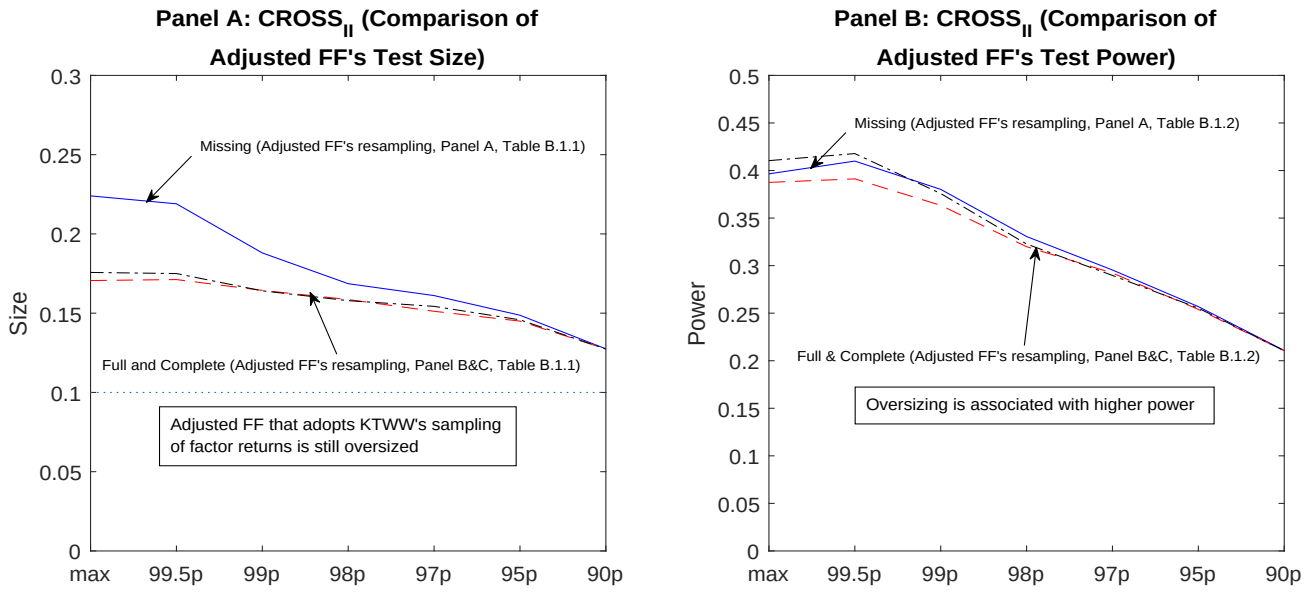


Figure 7: **Results: Adjusted FF's Test Size and Test Power, 1984–1988 (186 Funds).** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$. FF denotes Fama and French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

3.2.2 Test Power

We now choose non-zero levels of p (assumed fraction of outperforming funds) and IR (injected information ratio) to study test power (see Table 1 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 at 0.75 and p at 5%.

Figures 5, 6 and 7 also include a summary of test power and Table B.1.2 reports more detailed results, both corresponding to our baseline specification. In Panel A of Table B.1.2, FF's approach $CROSS_I^{mis}$, which corresponds to FF's sample selection, generates a 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^{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^{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). But their difference becomes smaller at lower percentiles.

The improved performance of the FF approach $CROSS_I$, when applied to the sample with the complete history of returns (i.e., $\mathcal{D}_{m,n}^{ful}$), can be explained by the results in Table B.1.1. 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 over sized, 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, a test power of 22.4% (i.e., the best case scenario for the sample $\mathcal{D}_{m,n}^{ful}$) still seems low. This low test power reflects more about the general difficulty in identifying outperforming funds for the mutual fund data than about a deficiency of the FF's approach. On the one hand, the close-to-optimal test size for $CROSS_I^{ful}$ in Table B.1.1 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.1.2, the average maximum t-statistic among truly performing funds is 2.94, which is not far away from the average maximum t-statistic, 2.65, among nonperforming funds. Moreover, the average maximum α for the former group is 11.07%, which is actually lower than the average maximum α for the latter.

Tables IA.1 to IA.8 in the Internet Appendix report our results under alternative values of IR and p. 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% significance) for the 99^{th} percentile.

Contrary to the perception that, for a given p of the fraction of outperforming funds, the $100(1\text{-}p)^{th}$ 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.1 in the Internet Appendix, the highest test power is found for the 99^{th} percentile, although $p = 10\%$ of funds are outperforming. In fact, test power for the 90^{th} percentile is only 37.5%, substantially lower than that for the 99^{th} percentile.

Overall, combining the evidence in Tables B.1.1 and B.1.2, we recommend the use of the FF approach with a complete history of returns (i.e., $CROSS_I^{ful}$). It has a near-optimal size and a 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^{th} percentile.

3.3 Results for the Full Sample

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

We first clarify how we obtain the 2,876 funds. Note our simulation design described in Section 3.1 cannot be directly applied because keeping funds that span the entire 1984–2018 period would leave us with very few funds. We adjust our simulation design as follows. First, driven by our results in Section 2.2.1, where $T = 60$ results in little distortion in the bootstrapped t-statistic distribution, we keep funds with at least 60 observations over the 1984–2018 period. This leaves us with 2,876 funds, which constitutes our $\mathcal{D}^{sub}$ for the 1984–2018 sample. Let the original sample of funds with at least 8 observations be $\mathcal{D}$, which has 4,007 funds.

Second, we follow the same procedure as described in Section 3.1 to inject alphas into funds in $\mathcal{D}^{sub}$ and obtain the corresponding $\mathcal{D}_m$ (see Figure 4).[11] 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 of 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 make sure the frequency distribution for the number of observations for $\mathcal{D}_m^{mis}$ is approximately the same as that for $\mathcal{D}$.[12] After obtaining $\mathcal{D}_m^{mis}$, we define $\mathcal{D}_m^{ful}$ as the subsample of funds in $\mathcal{D}_m^c$, where funds have at least 60 observations.

[11] One difference from the previous five-year setting is that we need to inject a different level of information ratio (IR). The reason is that with the same information ratio (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 an apples-to-apples comparison between our full sample results and subsample results. Our summary statistics reported in Table B.2.2 correspond well to those reported in Table B.1.2 and Table IB.2 in the Internet Appendix.

[12] For a fund with n_i observations in $\mathcal{D}_m^c$, we first keep its number of observations as n_i, if $n_i < 60$. Otherwise, we randomly generate a number (denoted as 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 of 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 keep 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.

Our results for the full sample are reported in Table B.2.1 and Table B.2.2 in the appendix. Figure 8 contains a summary for the FF approach. Figure B.2.1 and B.2.2 in the appendix contain the summary for IND_I and $CROSS_{II}$.

Figure B.2.1 and B.2.2 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.2.1), 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^{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^{th} and 99^{th}, have similar but lower test power compared to the five-year subsamples (e.g., the 1984–1988 subsample in Figure 6). The maximum statistic is somewhat under sized and therefore less powerful than alternative test statistics.

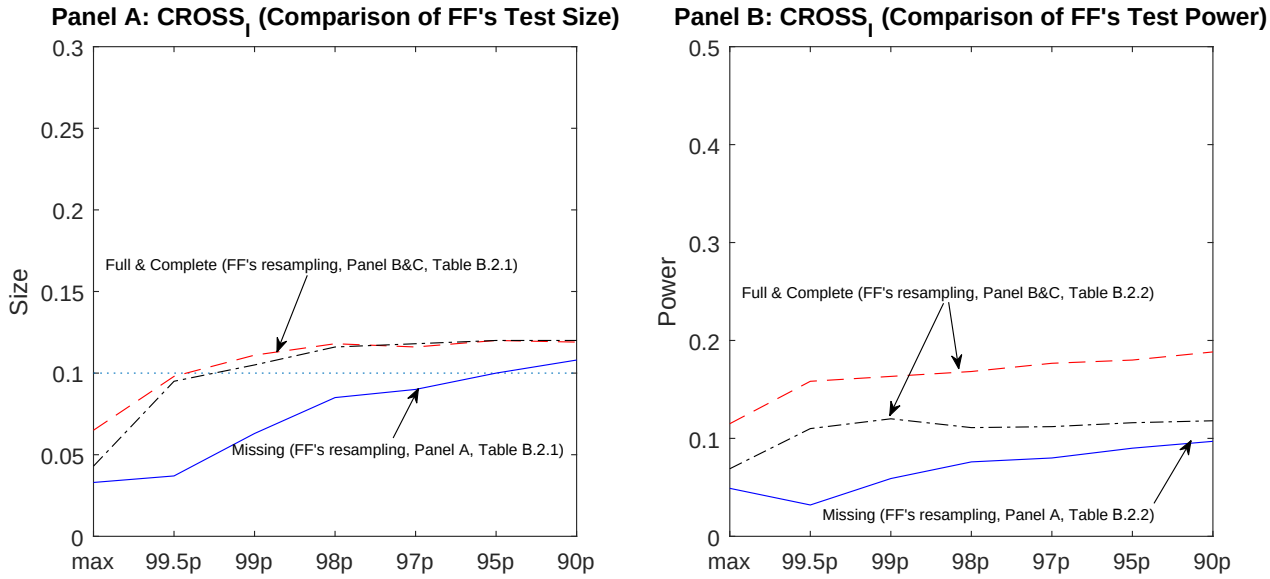


Figure 8: **Results: FF's Test Size and Test Power, full sample, 1984–2018, 2,876 funds.** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$.

3.4 Modifying FF: A Thresholding-FF Approach

While our strategy of only keeping funds with more than 60 observations helps mitigate the under-sampling issue of FF and enhance its test power, funds with fewer than 60 observations may represent an economically important set of funds (1,163 out 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, causing 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 of all, we keep all funds with a history of at least 12 monthly observations. While we can in principle keep all funds with at least eight observations, we believe 12 is a more reasonable cutoff given the increased instability of estimating t-statistic for funds with eight observations and four benchmark factors. Our thresholding-FF approach is then described as follows.

Before we perform the FF bootstrap, we do 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 only focus on months for which we observe fund i's returns and bootstrap 1,000 times (i.e., complete-sample bootstrap). Let the $25th$ and $75th$ percentile for the bootstrapped t-statistic distribution be $\hat{q}(25,i)$ and $\hat{q}(75,i)$. The bandwidth for t-statistic that we create for fund i is constructed as:

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

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, e.g., Tukey (1977)). As we shall see later, 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 do FF's missing-data bootstrap (after we subtract the in-sample alphas from all funds) and for bootstrap iteration b ($b = 1, \cdots, B = 1,000$), we discard fund i if its bootstrapped t-statistic falls out 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* for our data. In particular, we do a grid search within the set of $thres \in \{1.0, 1.5, 2.0, \cdots, 5.0\}$. For each value of *thres*, we simulate to find test size, i.e., 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 shows our results with a significance level of 10%.[13] Not surprisingly, test size is monotonically decreasing in *thres* because the higher *thres* is, 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).[14]

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 marked as "Full (FF's re-sampling)" in Figure 10), test size is well maintained at 10% for our thresholding-FF approach (which is also consistent with Figure 9). Meanwhile, test power is also higher,

[13] We choose 10% to be consistent with our previous figure displays. Our results are consistent across significance levels.

[14] Note 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. $\mathcal{D}^{sub}$ includes 2,876 funds.

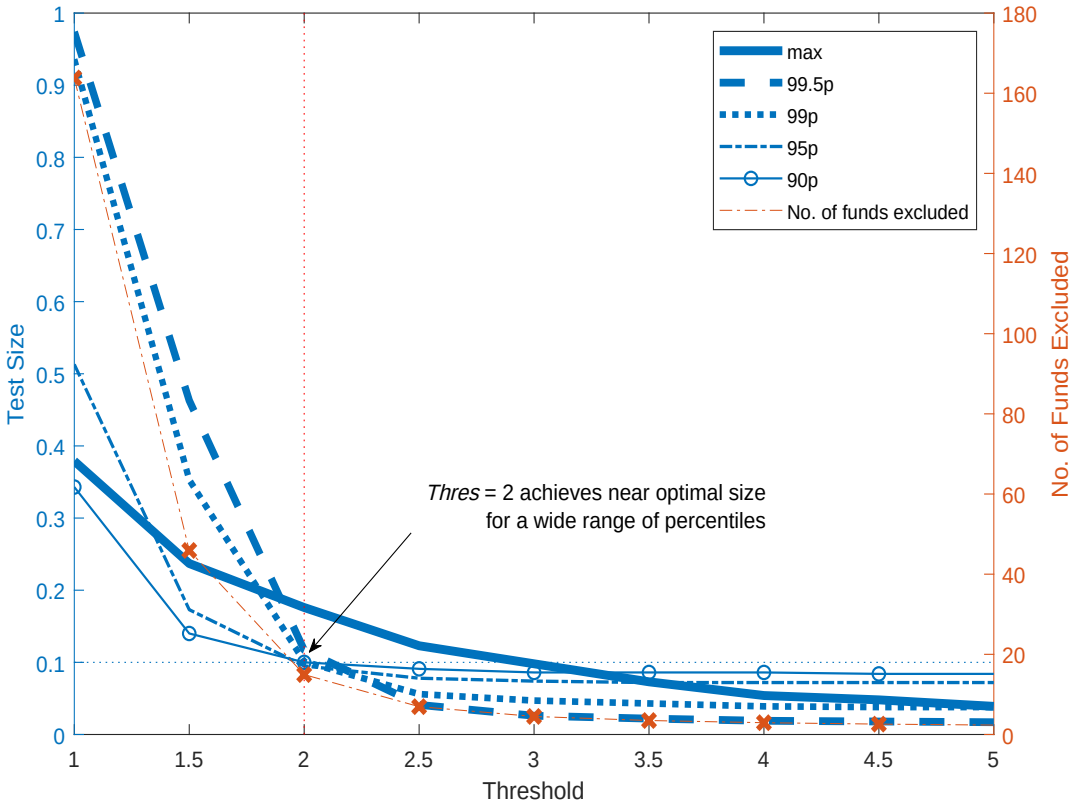


Figure 9: **Results: Simulated Test Size for the Thresholding-FF Approach, 2,876 funds.** We simulate to find the test size (left y-axis) for the thresholding-FF approach with a threshold parameter given by the x-axis. We also find the corresponding average number of funds dropped in the bootstrapped simulations (right y-axis).

especially for more extreme percentiles such as the $99.5th$ and $99th$ percentile. Overall, our thresholding-FF approach seems 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 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 by simply excluding 15 funds from the data.

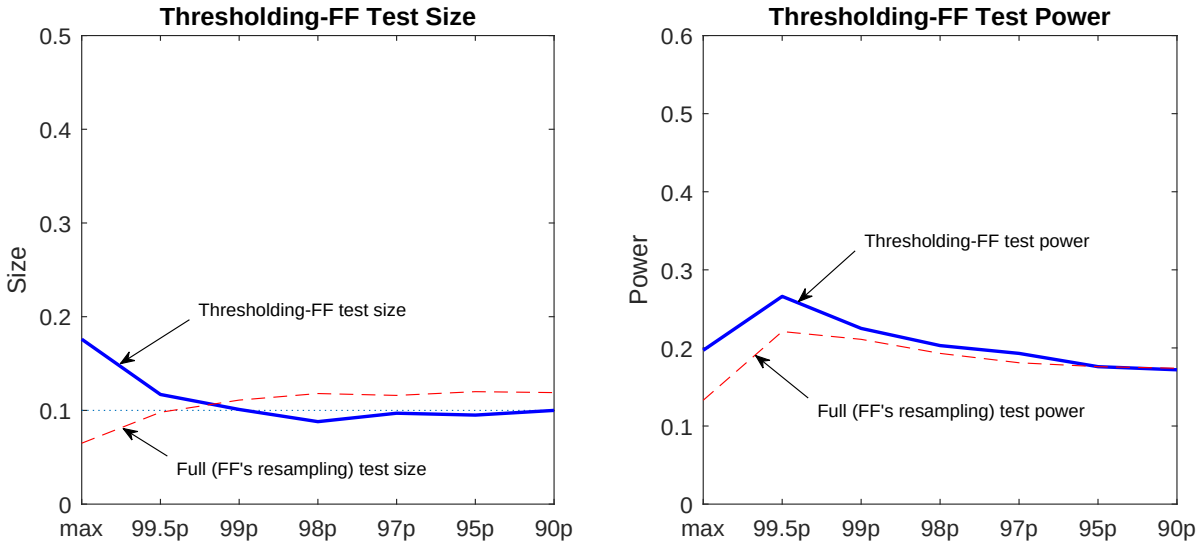


Figure 10: **Results: Simulated Test Size and Power for the Thresholding-FF Approach with** $thres = 2.0$**, 2,876 funds.** We report test size and test power at the 10% significance level for the thresholding-FF approach with the threshold parameter set at 2.0. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$.

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 datasets (e.g., hedge funds) that display a different level of signal-to-noise ratio in the performance metric or different dependence structure. Therefore, we recommend researchers conduct similar simulation studies to find the data-specific optimal value of $thres$.

4 Other Issues

Here we discuss several issues related to our simulation design.

4.1 Alternative Five-Year Subsamples

We also examine the 2014–2018 sample, which features a much larger number of funds (1,502) than for the 1984–1988 subsample (186). We report our results in Figure IB.1, Table IB.1 and Table IB.2 in the Internet Appendix.

Our findings are similar to the 1984-1988 subsample. Panel C of Figure IB.1 shows that the FF approach $CROSS_I$ overall performs well in test size for the full ($\mathcal{D}^{ful}_{m,n}$) or complete sample ($\mathcal{D}^{c}_{m,n}$). One exception is the maximum statistic, which appears to be oversized. FF do not consider the max 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 over-rejections. Nonetheless, starting from the 99.5^{th} percentile, less extreme percentiles do not seem to be subject to this concern.

4.2 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 non-zero-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.

5 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 two 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. 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 saying a fund outperforms) as well as the power (the probability of identifying a truly outperforming fund). In our simulation design, we exactly know which funds outperform, making it possible to measure these error rates.

There are two key differences in 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 8 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. That is, if we start with 23 fund observations, given that the cross-section is being resampled, we might draw fewer than 23 observations.

Our results suggest that the undersampling in 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 the 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 has the opposite problem. Our simulations show that KTWW substantially over-rejects. 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 with our thresholding approach. In our application, we can include funds with as few as 12 observations and achieve similar statistical performance to the approach that imposes a 60 observation minimum. In addition, our results may alter the interpretation of published papers that use the FF or KTWW bootstrap method.

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, 2020, Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? A Reply *Journal of Finance, Forthcoming.*

Beran, R., 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, P., 1997, Sieve bootstrap for time series, *Bernoulli 3, 123–148.*

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

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

Busse, Jeffrey A., Amit Goyal, 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.*

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

Chordia, T., A. Goyal, and A. Saretto, 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, R., and J. 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? *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning, 3753–3827.*

Giacomini, R., D. N. Politis, and H. White, 2013, A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory, 567–589.*

Hall, P., 1992, The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag.

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

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.* Available at https://ssrn.com/abstract=2358214.

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

Harvey, Campbell R., and Yan Liu, 2017, Luck vs. skill and factor selection, in *The Fama Portfolio, 250–260*, John Cochrane and Tobias J. Moskowitz, ed., Chicago: University of Chicago Press.

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.*

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, H., Lei Jiang, Xuan Leng, and Liang Peng, Bootstrap analysis of mutual fund performance, 2020, Working Paper.

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.*

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

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

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

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

MacKinnon, J. G., 2009, Bootstrap hypothesis testing, *Handbook of computational econometrics 183: 213.*

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

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

Tukey, J. W. 1977. Exploratory data analysis. Reading, PA. Addison-Wesley.

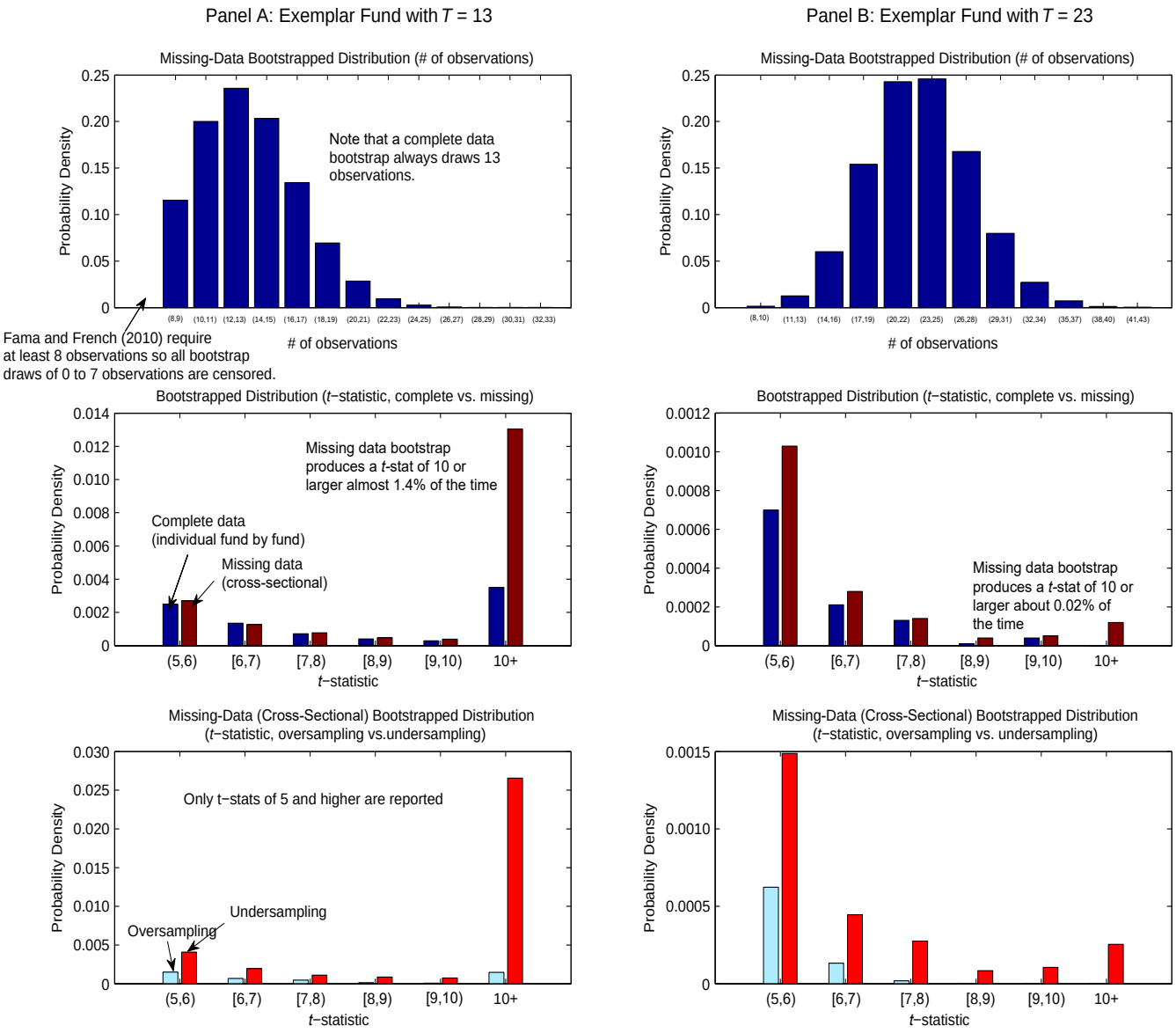
Yan, Xuemin, and Lingling Zheng, 2017, Fundamental analysis and the cross-section of stocks returns: A data-mining approach, *Review of Financial Studies 30, 1382–1423.*

A Illustration: Requiring Eight Observations (Including Non-Unique Observations)

We illustrate the undersampling issue with the stated approach (i.e., requiring eight observations, including non-unique observations) in Fama and French (2010) and compare with our results in Figure 1 and 2 (the actual approach used in Fama and French). This exercise is important because many researchers have implemented the stated approach. Our analysis will show that there are important differences between the stated and actual approaches for small samples.

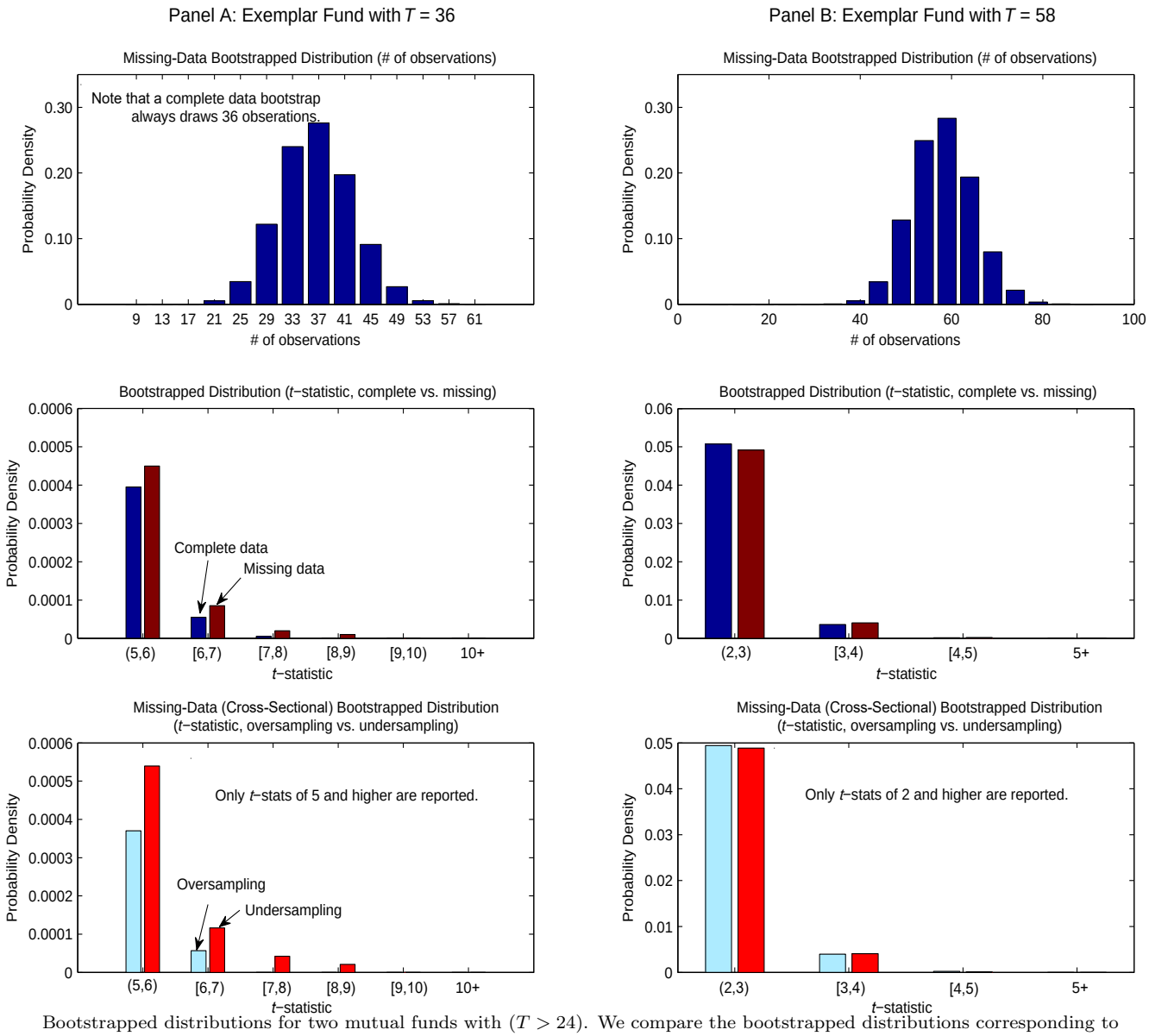
Several patterns emerge from the comparison. First, the difference is minor for a T above 36. The main differences stem from short-lived funds with a T below 36. Second, comparing Panel A in Figure 1 and Figure A.1.1, the censoring implied by the actual Fama-French approach (i.e., requiring eight unique observations), as shown in Figure 1, brings the missing-data bootstrapped distribution closer to its complete-data counterpart (than Figure A.1.1), although the missing-data bootstrap still leads to a higher chance 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.1.1, for funds with around two years of data, it is clear that either approach tilts the missing-data bootstrapped distribution more towards larger t-statistics than the complete-data bootstrapped distribution. It is also evident that undersampling is driving the results.

Figure A.1.1: **Bootstrapped Distributions for Two Mutual Funds with ($T \leq 24$)**



Bootstrapped distributions for two mutual funds with ($T \leq 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (individual fund by fund) and "missing data" bootstrap (Fama and French or cross-sectional bootstrap). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for 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. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For 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 stated censoring scheme that requires eight observations (including non-unique observations).

Figure A.1.2: **Bootstrapped Distributions for Two Mutual Funds with ($T > 24$)**



Bootstrapped distributions for two mutual funds with ($T > 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (individual fund by fund) and "missing data" bootstrap (Fama and French or cross-sectional approach). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for 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. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10 for panel A, and t-statistics with a value of two 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 stated censoring scheme that requires eight observations (including non-unique observations).

B Additional Results

B.1 Five-Year Subsample, 1984–1988

Table B.1.1: **Simulated Test Size for $T = 60$ (1984–1988)**

For all funds between 1984–1988 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 of $p = 0$ 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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
					CROSS^{mis}_{II}	*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^{mis}_{III}	*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

Continued on next page

Table B.1.1 – *Continued from previous page*

	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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	n.a.	n.a.	139.4	5.9	IND^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{III}	*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
Panel C: Sample is $\mathcal{D}^{c}_{m}$ (infeasible)													
# of funds	n.a.	n.a.	186.0	n.a.	IND^{c}_{I}	*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^{c}_{II}	*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^{c}_{I}	*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^{c}_{II}	*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^{c}_{III}	*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.1.2: **Simulated Test Power for $T = 60$ (1984–1988), information ratio $IR = 0.75$, and fraction of outperforming funds $p = 5\%$**

For all funds between 1984–1988 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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
					CROSS^{mis}_{II}	*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^{mis}_{III}	*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

Continued on next page

Table B.1.2 – *Continued from previous page*

	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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	6.7	1.3	132.7	5.8	IND^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{III}	*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
Panel C: Sample is $\mathcal{D}^{c}_{m}$ (infeasible)													
# of funds	9.0	n.a.	177.0	n.a.	IND^{c}_{I}	*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^{c}_{II}	*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^{c}_{I}	*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^{c}_{II}	*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^{c}_{III}	*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

B.2 Full Sample, 1984-2018

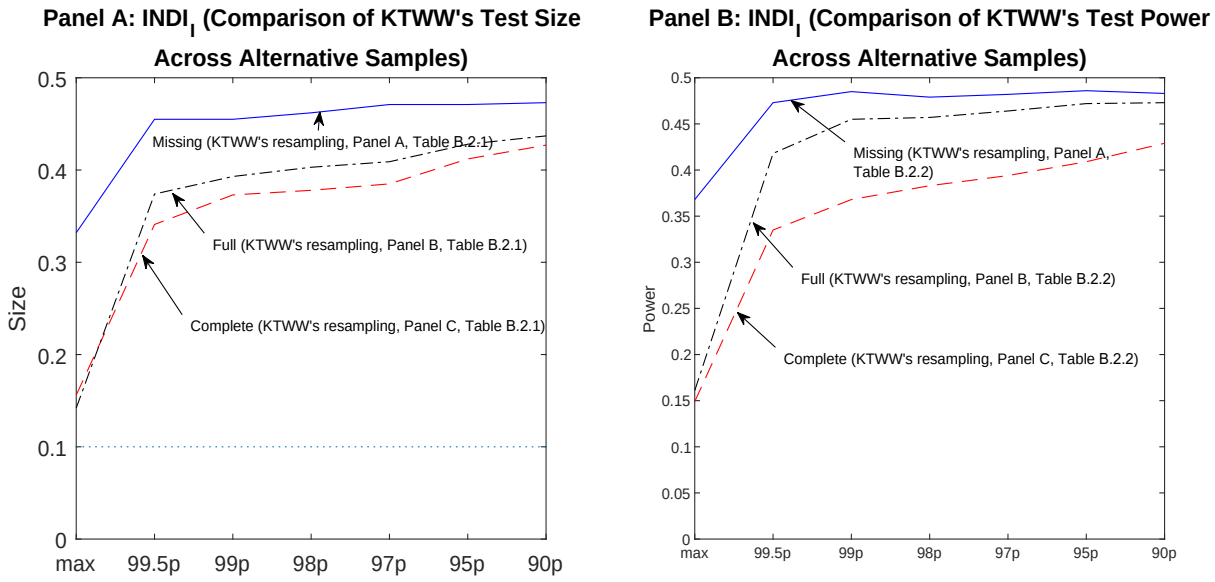


Figure B.2.1: **Results: KTWW's Test Size and Test Power, full sample, 1984–2018, 2,876 funds.** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$.

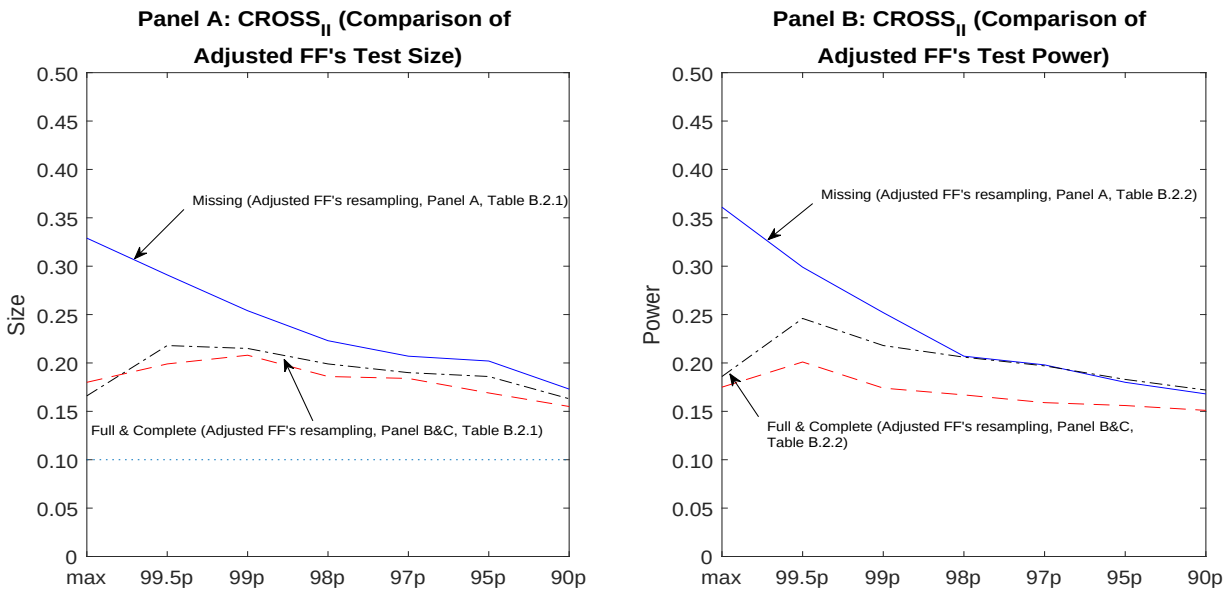


Figure B.2.2: **Results: Adjusted FF's Test Size and Test Power, full sample, 1984–2018, 2,876 funds.** We report test size and test power at the 10% significance level. Test size corresponds to setting $p = 0$. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$.

Table B.2.1: **Simulated Test Size for Full Sample (1984–2018)**

For all funds between 1984–2018 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 of $p = 0$ 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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
					CROSS^{mis}_I	*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^{mis}_{II}	*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^{mis}_{III}	*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

Continued on next page

Table B.2.1 – *Continued from previous page*

	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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with more than 60 observations)													
# of funds	n.a.	n.a.	2002.5	36.5	IND^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{III}	*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
Panel C: Sample is $\mathcal{D}^{c}_{m}$ (infeasible)													
# of funds	n.a.	n.a.	2876.0	n.a.	IND^{c}_{I}	*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^{c}_{II}	*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^{c}_{I}	*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^{c}_{II}	*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^{c}_{III}	*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.2.2: **Simulated Test Power for Full Sample (1984–2018), information ratio $IR = 0.75/\sqrt{7}$, and fraction of outperforming funds $p = 5\%$**

For all funds between 1984–2018 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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
					CROSS^{mis}_{II}	*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^{mis}_{III}	*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

Continued on next page

Table B.2.2 – *Continued from previous page*

	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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with more than 60 observations)													
# of funds	99.7	5.8	1900.9	35.2	IND^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{I}	*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^{ful}_{II}	*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^{ful}_{III}	*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
Panel C: Sample is $\mathcal{D}^{c}_{m}$ (infeasible)													
# of funds	143.0	n.a.	2733.0	n.a.	IND^{c}_{I}	*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^{c}_{II}	*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^{c}_{I}	*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^{c}_{II}	*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^{c}_{III}	*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

B.3 Fund Length Distribution

Table B.3.1: **Fund Length Distribution**

We summarize fund time-series length distributions across different sample periods. $p(10)$, $p(50)$, and $p(90)$ stand for 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.

Sample period	**Fund length in months**				
	min	$p(10)$	$p(50)$	$p(90)$	*max*
Panel A: 1984–1988 (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: 2014–2018 (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: 1984–2018 (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

Internet Appendix for "Luck versus Skill in the Cross-Section of Mutual Fund Returns: Reexamining the Evidence"

Campbell R. Harvey
Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu*
Purdue University, West Lafayette, IN 47906 USA

Current version: October 17, 2021

* Current Version: October 17, 2021. First posted on SSRN: June 9, 2020.

1 Summary

Appendix A provides additional results for the 1984–1988 sample with alternative specifications of IR (injected information ratio) and p (fraction of outperforming funds).

Table IA.1 has $IR = 1.0$ and $p = 10\%$.
Table IA.2 has $IR = 1.0$ and $p = 5\%$.
Table IA.3 has $IR = 1.0$ and $p = 2.5\%$.
Table IA.4 has $IR = 0.75$ and $p = 10\%$.
Table IA.5 has $IR = 0.75$ and $p = 2.5\%$.
Table IA.6 has $IR = 0.5$ and $p = 10\%$.
Table IA.7 has $IR = 0.5$ and $p = 5\%$.
Table IA.8 has $IR = 0.5$ and $p = 2.5\%$.

Appendix B provides results for the 2014–2018 sample.

A 1984–1988, Alternative Specifications

Table IA.1: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 1.0$, and fraction of outperforming funds $p = 10\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 1.0 and the fraction of outperforming funds of 10%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 1.0 to $p = 10\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.1 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	18.0	n.a.	168.0	n.a.	IND_{I}^{mis}	*1%*	0.225	0.386	0.599	0.690	0.703	0.693	0.589
avg. t-stat	1.92	0.39	−0.04	0.26		*5%*	0.565	0.692	0.797	0.836	0.831	0.808	0.700
avg. α(%)	6.11	1.56	−0.07	0.86		*10%*	0.724	0.808	0.869	0.889	0.882	0.855	0.754
max t-stat	4.04	0.84	3.01	0.93									
max α *(%)*	20.96	11.45	21.11	12.99									
					IND_{II}^{mis}	*1%*	0.227	0.389	0.604	0.689	0.705	0.693	0.589
						5%	0.569	0.695	0.797	0.836	0.832	0.807	0.700
						10%	0.727	0.810	0.870	0.889	0.883	0.855	0.754
					CROSS_{I}^{mis}	*1%*	0.001	0.002	0.047	0.082	0.075	0.060	0.041
						5%	0.041	0.066	0.269	0.330	0.312	0.268	0.192
						10%	0.157	0.218	0.484	0.539	0.521	0.456	0.346
					CROSS_{II}^{mis}	*1%*	0.238	0.326	0.389	0.343	0.288	0.202	0.105
						5%	0.598	0.671	0.709	0.668	0.614	0.508	0.324
						10%	0.766	0.815	0.835	0.813	0.771	0.686	0.494
					CROSS_{III}^{mis}	*1%*	0.243	0.331	0.392	0.344	0.287	0.206	0.104
						5%	0.602	0.674	0.712	0.670	0.615	0.509	0.326
						10%	0.767	0.817	0.837	0.813	0.771	0.686	0.495

Continued on next page

Table IA.1 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	13.5	1.8	126.0	5.6	IND^{ful}_{I}	*1%*	0.341	0.420	0.568	0.675	0.692	0.671	0.569
avg. t-stat	2.07	0.42	−0.04	0.29		*5%*	0.620	0.680	0.776	0.821	0.823	0.792	0.682
avg. α(%)	6.13	1.62	−0.05	0.85		*10%*	0.754	0.792	0.856	0.877	0.876	0.842	0.739
max t-stat	3.92	0.80	2.64	0.67									
max α *(%)*	17.84	9.96	13.56	7.61									
					IND^{ful}_{II}	*1%*	0.346	0.428	0.574	0.675	0.694	0.671	0.569
						5%	0.624	0.683	0.777	0.821	0.822	0.792	0.681
						10%	0.756	0.795	0.856	0.878	0.875	0.842	0.739
					CROSS^{ful}_{I}	*1%*	0.118	0.126	0.133	0.123	0.106	0.081	0.051
						5%	0.389	0.406	0.428	0.411	0.381	0.313	0.216
						10%	0.587	0.608	0.632	0.620	0.586	0.510	0.377
					CROSS^{ful}_{II}	*1%*	0.356	0.385	0.391	0.354	0.305	0.220	0.114
						5%	0.662	0.687	0.703	0.672	0.624	0.521	0.333
						10%	0.802	0.822	0.833	0.811	0.776	0.689	0.499
					CROSS^{ful}_{III}	*1%*	0.363	0.392	0.396	0.357	0.307	0.220	0.116
						5%	0.667	0.692	0.707	0.675	0.627	0.521	0.334
						10%	0.804	0.824	0.834	0.811	0.776	0.690	0.501

Continued on next page

Table IA.1 – *Continued from previous page*

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	18.0	n.a.	168.0	n.a.	IND_I^c	*1%*	0.361	0.547	0.687	0.760	0.765	0.741	0.620
avg. t-stat	2.07	0.39	−0.04	0.28		*5%*	0.651	0.773	0.845	0.875	0.869	0.840	0.718
avg. α(%)	6.13	1.62	−0.05	0.83		*10%*	0.781	0.857	0.903	0.918	0.909	0.881	0.769
max t-stat	4.08	0.79	2.74	0.67									
max α *(%)*	19.72	10.45	14.78	7.95									
					IND_{II}^c	*1%*	0.366	0.552	0.691	0.760	0.764	0.741	0.621
						5%	0.656	0.776	0.846	0.875	0.868	0.840	0.718
						10%	0.785	0.859	0.904	0.918	0.909	0.882	0.768
					CROSS_I^c	*1%*	0.119	0.134	0.139	0.121	0.104	0.077	0.050
						5%	0.398	0.440	0.452	0.424	0.384	0.313	0.212
						10%	0.603	0.648	0.663	0.643	0.604	0.521	0.375
					CROSS_{II}^c	*1%*	0.383	0.437	0.429	0.372	0.311	0.219	0.113
						5%	0.697	0.750	0.747	0.708	0.646	0.534	0.333
						10%	0.831	0.866	0.870	0.841	0.802	0.712	0.505
					CROSS_{III}^c	*1%*	0.391	0.446	0.433	0.374	0.313	0.217	0.111
						5%	0.701	0.752	0.749	0.708	0.650	0.534	0.334
						10%	0.834	0.868	0.871	0.843	0.801	0.712	0.506

Table IA.2: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 1.0$, and fraction of outperforming funds $p = 5\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 1.0 and the fraction of outperforming funds of 5%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 1.0 to $p = 5\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.2 – *Continued from previous page*

	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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	9.0	n.a.	177.0	n.a.	IND_{I}^{mis}	*1%*	0.141	0.228	0.346	0.379	0.376	0.360	0.331
avg. t-stat	1.91	0.48	−0.05	0.26		*5%*	0.395	0.479	0.548	0.553	0.529	0.491	0.437
avg. α(%)	6.01	2.15	−0.09	0.86		*10%*	0.543	0.608	0.652	0.644	0.612	0.563	0.497
max t-stat	3.61	0.84	3.02	0.95									
max α *(%)*	15.81	9.43	21.28	12.67									
					IND_{II}^{mis}	*1%*	0.143	0.229	0.345	0.379	0.376	0.360	0.331
						5%	0.397	0.481	0.551	0.555	0.529	0.492	0.435
						10%	0.546	0.610	0.655	0.644	0.613	0.563	0.496
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.017	0.024	0.021	0.020	0.016
						5%	0.024	0.033	0.125	0.132	0.116	0.100	0.084
						10%	0.099	0.123	0.261	0.262	0.235	0.198	0.167
					CROSS_{II}^{mis}	*1%*	0.153	0.188	0.186	0.129	0.097	0.066	0.039
						5%	0.423	0.459	0.443	0.357	0.293	0.218	0.149
						10%	0.587	0.618	0.605	0.522	0.447	0.352	0.256
					CROSS_{III}^{mis}	*1%*	0.153	0.190	0.190	0.131	0.098	0.066	0.039
						5%	0.427	0.463	0.446	0.359	0.295	0.219	0.150
						10%	0.590	0.620	0.605	0.524	0.448	0.354	0.256

Continued on next page

Table IA.2 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	6.7	1.3	132.7	5.8	IND^{ful}_{I}	*1%*	0.204	0.246	0.320	0.364	0.359	0.338	0.309
avg. t-stat	2.07	0.52	−0.04	0.28		*5%*	0.423	0.463	0.523	0.538	0.513	0.466	0.415
avg. α(%)	6.06	2.26	−0.07	0.85		*10%*	0.558	0.589	0.629	0.623	0.593	0.536	0.476
max t-stat	3.49	0.84	2.65	0.67									
max α *(%)*	13.49	8.23	13.64	7.58									
					IND^{ful}_{II}	*1%*	0.207	0.251	0.324	0.365	0.361	0.338	0.311
						5%	0.426	0.467	0.525	0.538	0.513	0.466	0.415
						10%	0.560	0.593	0.630	0.625	0.594	0.535	0.476
					CROSS^{ful}_{I}	*1%*	0.064	0.063	0.053	0.040	0.033	0.028	0.021
						5%	0.244	0.245	0.226	0.182	0.150	0.118	0.097
						10%	0.400	0.405	0.385	0.330	0.284	0.226	0.184
					CROSS^{ful}_{II}	*1%*	0.216	0.223	0.194	0.138	0.101	0.071	0.043
						5%	0.460	0.472	0.441	0.363	0.298	0.220	0.152
						10%	0.614	0.625	0.597	0.523	0.449	0.352	0.255
					CROSS^{ful}_{III}	*1%*	0.219	0.228	0.197	0.139	0.102	0.071	0.043
						5%	0.463	0.475	0.444	0.363	0.299	0.222	0.153
						10%	0.617	0.627	0.598	0.525	0.449	0.354	0.256

Continued on next page

Table IA.2 – *Continued from previous page*

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%*	0.221	0.329	0.407	0.424	0.406	0.378	0.349
avg. t-stat	2.06	0.47	−0.04	0.28		*5%*	0.454	0.547	0.599	0.589	0.552	0.498	0.447
avg. α(%)	6.04	1.97	−0.07	0.83		*10%*	0.590	0.663	0.691	0.673	0.629	0.568	0.501
max t-stat	3.67	0.81	2.75	0.67									
max α *(%)*	14.95	8.67	14.88	7.93									
					IND_{II}^c	*1%*	0.225	0.334	0.410	0.426	0.407	0.379	0.350
						5%	0.458	0.548	0.600	0.588	0.552	0.498	0.447
						10%	0.593	0.665	0.692	0.673	0.628	0.567	0.502
					CROSS_I^c	*1%*	0.065	0.063	0.052	0.038	0.031	0.026	0.021
						5%	0.255	0.257	0.229	0.176	0.143	0.115	0.094
						10%	0.415	0.428	0.395	0.326	0.276	0.222	0.183
					CROSS_{II}^c	*1%*	0.236	0.247	0.199	0.133	0.097	0.069	0.043
						5%	0.497	0.519	0.472	0.368	0.295	0.217	0.152
						10%	0.655	0.678	0.634	0.537	0.452	0.350	0.258
					CROSS_{III}^c	*1%*	0.240	0.251	0.203	0.135	0.099	0.068	0.0425
						5%	0.502	0.522	0.474	0.369	0.296	0.218	0.151
						10%	0.657	0.680	0.635	0.538	0.454	0.350	0.259

Table IA.3: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 1.0$, and fraction of outperforming funds $p = 2.5\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 1.0 and the fraction of outperforming funds of 2.5%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 1.0 to $p = 2.5\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.3 – *Continued from previous page*

	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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	4.0	n.a.	182.0	n.a.	IND_{I}^{mis}	*1%*	0.095	0.131	0.178	0.199	0.210	0.231	0.241
avg. t-stat	1.92	0.65	−0.04	0.26		*5%*	0.261	0.308	0.333	0.330	0.326	0.329	0.328
avg. α(%)	6.19	3.11	−0.07	0.86		*10%*	0.383	0.417	0.429	0.407	0.398	0.388	0.378
max t-stat	3.09	0.91	3.05	0.96									
max α *(%)*	12.22	8.31	21.70	13.34									
					IND_{II}^{mis}	*1%*	0.095	0.131	0.181	0.200	0.210	0.231	0.243
						5%	0.263	0.310	0.335	0.328	0.327	0.329	0.328
						10%	0.387	0.421	0.430	0.408	0.398	0.389	0.377
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.007	0.011	0.012	0.012	0.011
						5%	0.018	0.020	0.054	0.060	0.062	0.059	0.056
						10%	0.065	0.069	0.131	0.133	0.128	0.123	0.117
					CROSS_{II}^{mis}	*1%*	0.100	0.107	0.087	0.056	0.046	0.037	0.026
						5%	0.284	0.290	0.247	0.184	0.157	0.130	0.098
						10%	0.424	0.427	0.383	0.304	0.261	0.224	0.181
					CROSS_{III}^{mis}	*1%*	0.100	0.107	0.087	0.057	0.046	0.037	0.026
						5%	0.288	0.294	0.248	0.185	0.158	0.130	0.099
						10%	0.427	0.429	0.384	0.306	0.261	0.224	0.181

Continued on next page

Table IA.3 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	3.0	0.8	136.4	5.8	IND^{ful}_{I}	*1%*	0.111	0.130	0.159	0.181	0.191	0.209	0.223
avg. t-stat	2.07	0.73	−0.04	0.29		*5%*	0.265	0.284	0.304	0.303	0.303	0.305	0.310
avg. α(%)	6.21	3.39	−0.05	0.85		*10%*	0.371	0.389	0.393	0.380	0.370	0.365	0.359
max t-stat	2.93	0.96	2.66	0.67									
max α *(%)*	10.24	7.17	13.86	7.79									
					IND^{ful}_{II}	*1%*	0.111	0.132	0.160	0.181	0.192	0.209	0.2236
						5%	0.267	0.287	0.305	0.303	0.302	0.306	0.310
						10%	0.374	0.391	0.396	0.381	0.371	0.365	0.359
					CROSS^{ful}_{I}	*1%*	0.035	0.032	0.022	0.016	0.016	0.016	0.013
						5%	0.138	0.132	0.106	0.083	0.075	0.070	0.062
						10%	0.250	0.245	0.208	0.167	0.149	0.138	0.127
					CROSS^{ful}_{II}	*1%*	0.117	0.115	0.083	0.054	0.046	0.037	0.027
						5%	0.294	0.292	0.240	0.179	0.151	0.128	0.101
						10%	0.424	0.423	0.367	0.293	0.253	0.219	0.179
					CROSS^{ful}_{III}	*1%*	0.120	0.117	0.085	0.056	0.046	0.037	0.026
						5%	0.296	0.293	0.241	0.181	0.152	0.130	0.101
						10%	0.425	0.426	0.368	0.293	0.253	0.220	0.178

Continued on next page

Table IA.3 – *Continued from previous page*

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	4.0	n.a.	182.0	n.a.	IND_I^c	*1%*	0.125	0.175	0.203	0.213	0.222	0.241	0.256
avg. t-stat	2.07	0.63	−0.04	0.28		*5%*	0.289	0.338	0.346	0.331	0.331	0.333	0.336
avg. α(%)	6.21	2.78	−0.05	0.83		*10%*	0.403	0.438	0.433	0.401	0.393	0.389	0.383
max t-stat	3.18	0.88	2.76	0.67									
max α *(%)*	11.69	7.60	15.05	8.08									
					IND_{II}^c	*1%*	0.128	0.178	0.204	0.214	0.222	0.241	0.255
						5%	0.293	0.341	0.347	0.331	0.331	0.333	0.335
						10%	0.406	0.441	0.433	0.401	0.393	0.388	0.383
					CROSS_I^c	*1%*	0.036	0.030	0.020	0.016	0.016	0.016	0.013
						5%	0.148	0.132	0.101	0.079	0.075	0.070	0.063
						10%	0.266	0.251	0.205	0.161	0.148	0.138	0.127
					CROSS_{II}^c	*1%*	0.134	0.124	0.081	0.053	0.045	0.038	0.026
						5%	0.325	0.315	0.247	0.177	0.152	0.129	0.100
						10%	0.460	0.452	0.379	0.290	0.252	0.219	0.180
					CROSS_{III}^c	*1%*	0.137	0.126	0.082	0.054	0.045	0.037	0.027
						5%	0.328	0.318	0.249	0.180	0.153	0.129	0.100
						10%	0.464	0.456	0.382	0.292	0.254	0.219	0.181

Table IA.4: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 0.75$, and fraction of outperforming funds $p = 10\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 0.75 and the fraction of outperforming funds of 10%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 to $p = 10\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.4 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	18.0	n.a.	168.0	n.a.	IND_{I}^{mis}	*1%*	0.115	0.186	0.307	0.381	0.410	0.436	0.427
avg. t-stat	1.43	0.37	−0.04	0.26		*5%*	0.331	0.418	0.509	0.558	0.569	0.573	0.541
avg. α(%)	4.56	1.47	−0.07	0.86		*10%*	0.479	0.551	0.616	0.647	0.651	0.642	0.600
max t-stat	3.49	0.81	3.01	0.93									
max α *(%)*	17.57	10.13	21.11	13.01									
					IND_{II}^{mis}	*1%*	0.115	0.185	0.310	0.383	0.410	0.435	0.428
						5%	0.333	0.421	0.512	0.557	0.568	0.572	0.541
						10%	0.482	0.554	0.616	0.647	0.652	0.642	0.601
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.014	0.025	0.027	0.025	0.022
						5%	0.020	0.028	0.108	0.139	0.135	0.128	0.116
						10%	0.080	0.102	0.231	0.267	0.262	0.249	0.226
					CROSS_{II}^{mis}	*1%*	0.120	0.153	0.164	0.133	0.114	0.086	0.057
						5%	0.360	0.397	0.403	0.362	0.325	0.270	0.206
						10%	0.527	0.561	0.566	0.525	0.487	0.427	0.337
					CROSS_{III}^{mis}	*1%*	0.122	0.155	0.166	0.135	0.113	0.088	0.058
						5%	0.362	0.399	0.404	0.362	0.327	0.271	0.204
						10%	0.531	0.563	0.566	0.526	0.486	0.427	0.337

Continued on next page

Table IA.4 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	13.5	1.8	126.0	5.6	IND^{ful}_{I}	*1%*	0.150	0.190	0.276	0.362	0.388	0.405	0.405
avg. t-stat	1.54	0.41	−0.04	0.29		*5%*	0.355	0.400	0.475	0.528	0.545	0.548	0.519
avg. α(%)	4.59	1.48	−0.05	0.85		*10%*	0.486	0.518	0.586	0.621	0.629	0.621	0.581
max t-stat	3.36	0.77	2.64	0.67									
max α *(%)*	14.63	8.53	13.56	7.60									
					IND^{ful}_{II}	*1%*	0.153	0.192	0.278	0.365	0.388	0.407	0.404
						5%	0.358	0.403	0.478	0.529	0.546	0.548	0.520
						10%	0.489	0.521	0.588	0.621	0.629	0.622	0.582
					CROSS^{ful}_{I}	*1%*	0.044	0.044	0.045	0.039	0.036	0.032	0.029
						5%	0.187	0.190	0.193	0.180	0.167	0.148	0.131
						10%	0.333	0.342	0.340	0.324	0.305	0.277	0.246
					CROSS^{ful}_{II}	*1%*	0.162	0.170	0.158	0.134	0.114	0.089	0.062
						5%	0.394	0.408	0.395	0.362	0.323	0.271	0.207
						10%	0.544	0.558	0.554	0.516	0.480	0.423	0.340
					CROSS^{ful}_{III}	*1%*	0.164	0.173	0.160	0.135	0.115	0.089	0.062
						5%	0.397	0.412	0.398	0.362	0.326	0.274	0.208
						10%	0.546	0.558	0.556	0.516	0.481	0.425	0.341

Continued on next page

Table IA.4 – *Continued from previous page*

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	18.0	n.a.	168.0	n.a.	IND_I^c	*1%*	0.160	0.261	0.355	0.423	0.445	0.460	0.452
avg. t-stat	1.54	0.38	−0.04	0.28		*5%*	0.372	0.468	0.540	0.582	0.590	0.587	0.558
avg. α(%)	4.59	1.31	−0.05	0.83		*10%*	0.507	0.582	0.637	0.663	0.666	0.654	0.612
max t-stat	3.52	0.76	2.74	0.67									
max α *(%)*	16.18	8.96	14.78	7.94									
					IND_{II}^c	*1%*	0.164	0.264	0.360	0.426	0.443	0.461	0.452
						5%	0.375	0.473	0.541	0.583	0.591	0.588	0.557
						10%	0.512	0.586	0.640	0.664	0.667	0.654	0.612
					CROSS_I^c	*1%*	0.044	0.046	0.044	0.039	0.036	0.032	0.028
						5%	0.191	0.198	0.195	0.178	0.164	0.149	0.130
						10%	0.339	0.354	0.350	0.328	0.307	0.278	0.246
					CROSS_{II}^c	*1%*	0.174	0.189	0.171	0.134	0.113	0.089	0.062
						5%	0.415	0.440	0.417	0.371	0.330	0.276	0.210
						10%	0.573	0.597	0.576	0.533	0.490	0.428	0.340
					CROSS_{III}^c	*1%*	0.174	0.191	0.172	0.135	0.112	0.089	0.062
						5%	0.420	0.443	0.419	0.372	0.329	0.276	0.210
						10%	0.575	0.601	0.579	0.533	0.489	0.430	0.340

Table IA.5: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 0.75$, and fraction of outperforming funds $p = 2.5\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 0.75 and the fraction of outperforming funds of 2.5%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 to $p = 2.5\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.5 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	4.0	n.a.	182.0	n.a.	IND_{I}^{mis}	*1%*	0.063	0.085	0.117	0.151	0.173	0.201	0.223
avg. t-stat	1.43	0.63	−0.04	0.26		*5%*	0.180	0.210	0.234	0.261	0.275	0.292	0.307
avg. α(%)	4.63	2.88	−0.07	0.86		*10%*	0.274	0.298	0.317	0.328	0.338	0.348	0.355
max t-stat	2.57	0.88	3.05	0.96									
max α *(%)*	9.98	7.33	21.70	13.31									
					IND_{II}^{mis}	*1%*	0.063	0.086	0.118	0.152	0.173	0.201	0.223
						5%	0.180	0.211	0.236	0.260	0.276	0.292	0.307
						10%	0.275	0.300	0.318	0.329	0.338	0.348	0.355
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.005	0.008	0.010	0.010	0.010
						5%	0.013	0.013	0.035	0.046	0.050	0.051	0.051
						10%	0.044	0.045	0.086	0.102	0.106	0.107	0.106
					CROSS_{II}^{mis}	*1%*	0.067	0.069	0.053	0.040	0.036	0.030	0.023
						5%	0.196	0.196	0.167	0.140	0.126	0.111	0.090
						10%	0.307	0.307	0.275	0.239	0.217	0.195	0.164
					CROSS_{III}^{mis}	*1%*	0.066	0.070	0.054	0.041	0.037	0.031	0.023
						5%	0.197	0.198	0.168	0.141	0.127	0.112	0.090
						10%	0.309	0.309	0.277	0.240	0.217	0.195	0.165

Continued on next page

Table IA.5 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	3.0	0.8	136.4	5.8	IND^{ful}_{I}	*1%*	0.056	0.069	0.097	0.133	0.156	0.182	0.205
avg. t-stat	1.54	0.72	−0.04	0.29		*5%*	0.156	0.172	0.202	0.231	0.250	0.270	0.291
avg. α(%)	4.65	3.05	−0.05	0.84		*10%*	0.236	0.252	0.279	0.297	0.314	0.325	0.338
max t-stat	2.39	0.94	2.66	0.67									
max α *(%)*	8.19	6.19	13.86	7.79									
					IND^{ful}_{II}	*1%*	0.058	0.071	0.098	0.134	0.157	0.182	0.205
						5%	0.158	0.174	0.202	0.231	0.252	0.271	0.291
						10%	0.239	0.254	0.279	0.299	0.313	0.325	0.338
					CROSS^{ful}_{I}	*1%*	0.016	0.015	0.014	0.012	0.013	0.013	0.012
						5%	0.077	0.073	0.066	0.062	0.061	0.060	0.057
						10%	0.152	0.190	0.139	0.126	0.123	0.120	0.116
					CROSS^{ful}_{II}	*1%*	0.061	0.061	0.047	0.039	0.035	0.031	0.024
						5%	0.180	0.178	0.155	0.132	0.122	0.110	0.091
						10%	0.281	0.280	0.254	0.221	0.207	0.191	0.163
					CROSS^{ful}_{III}	*1%*	0.062	0.061	0.048	0.040	0.035	0.031	0.024
						5%	0.180	0.179	0.156	0.134	0.123	0.111	0.091
						10%	0.281	0.283	0.255	0.221	0.208	0.191	0.162

Continued on next page

Table IA.5 – *Continued from previous page*

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	4.0	n.a.	182.0	n.a.	IND_I^c	*1%*	0.061	0.094	0.127	0.162	0.183	0.211	0.236
avg. t-stat	1.54	0.62	−0.04	0.28		*5%*	0.167	0.201	0.233	0.261	0.276	0.297	0.314
avg. α(%)	4.65	2.53	−0.05	0.83		*10%*	0.251	0.283	0.303	0.321	0.335	0.350	0.359
max t-stat	2.64	0.786	2.76	0.67									
max α *(%)*	9.42	6.54	15.05	8.08									
					IND_{II}^c	*1%*	0.063	0.095	0.128	0.163	0.183	0.212	0.236
						5%	0.169	0.205	0.235	0.260	0.276	0.297	0.314
						10%	0.253	0.286	0.304	0.321	0.335	0.350	0.359
					CROSS_I^c	*1%*	0.016	0.014	0.013	0.013	0.013	0.014	0.0113
						5%	0.079	0.071	0.065	0.060	0.062	0.061	0.057
						10%	0.156	0.147	0.136	0.126	0.124	0.121	0.117
					CROSS_{II}^c	*1%*	0.067	0.062	0.048	0.040	0.036	0.031	0.023
						5%	0.195	0.187	0.158	0.135	0.124	0.111	0.090
						10%	0.298	0.295	0.259	0.226	0.209	0.191	0.164
					CROSS_{III}^c	*1%*	0.069	0.063	0.049	0.039	0.036	0.031	0.024
						5%	0.197	0.188	0.160	0.136	0.125	0.110	0.090
						10%	0.301	0.298	0.262	0.227	0.210	0.192	0.165

Table IA.6: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 0.5$, and fraction of outperforming funds $p = 10\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 0.5 and the fraction of outperforming funds of 10%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 0.5 to $p = 10\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.6 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	18.0	n.a.	168.0	n.a.	IND_I^{mis}	*1%*	0.067	0.098	0.156	0.208	0.238	0.277	0.303
avg. t-stat	0.94	0.36	−0.04	0.26		*5%*	0.201	0.244	0.295	0.338	0.358	0.387	0.403
avg. α(%)	3.01	1.39	−0.07	0.86		*10%*	0.304	0.344	0.386	0.418	0.432	0.454	0.461
max t-stat	2.96	0.79	3.01	0.93									
max α *(%)*	14.35	8.88	21.11	13.01									
					IND_{II}^{mis}	*1%*	0.067	0.099	0.156	0.209	0.238	0.277	0.303
						5%	0.201	0.247	0.297	0.338	0.360	0.388	0.402
						10%	0.305	0.346	0.388	0.419	0.433	0.454	0.460
					CROSS_I^{mis}	*1%*	0.001	0.001	0.005	0.012	0.013	0.014	0.013
						5%	0.013	0.014	0.049	0.067	0.071	0.074	0.074
						10%	0.046	0.052	0.116	0.142	0.146	0.149	0.150
					CROSS_{II}^{mis}	*1%*	0.072	0.080	0.073	0.061	0.055	0.047	0.032
						5%	0.219	0.231	0.218	0.193	0.179	0.158	0.133
						10%	0.342	0.353	0.341	0.311	0.291	0.269	0.233
					CROSS_{III}^{mis}	*1%*	0.072	0.081	0.074	0.062	0.054	0.046	0.033
						5%	0.221	0.232	0.218	0.195	0.179	0.157	0.133
						10%	0.346	0.356	0.342	0.312	0.291	0.268	0.233

Continued on next page

Table IA.6 – *Continued from previous page*

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 B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only funds with a complete history of returns)													
# of funds	13.5	1.8	126.0	5.6	IND_{I}^{ful}	*1%*	0.065	0.084	0.128	0.186	0.215	0.247	0.286
avg. t-stat	1.02	0.40	−0.04	0.29		*5%*	0.182	0.206	0.259	0.309	0.334	0.362	0.386
avg. α(%)	3.04	1.36	−0.05	0.85		*10%*	0.273	0.294	0.345	0.386	0.406	0.428	0.443
max t-stat	2.81	0.75	2.64	0.67									
max α *(%)*	11.55	7.10	13.56	7.61									
					IND_{II}^{ful}	*1%*	0.066	0.086	0.130	0.188	0.218	0.248	0.284
						5%	0.184	0.208	0.260	0.309	0.334	0.363	0.385
						10%	0.275	0.296	0.347	0.387	0.406	0.429	0.443
					CROSS_{I}^{ful}	*1%*	0.016	0.017	0.018	0.018	0.018	0.018	0.018
						5%	0.091	0.090	0.089	0.087	0.086	0.084	0.083
						10%	0.176	0.177	0.177	0.172	0.171	0.167	0.164
					CROSS_{II}^{ful}	*1%*	0.071	0.072	0.066	0.059	0.052	0.045	0.035
						5%	0.208	0.212	0.202	0.185	0.173	0.155	0.133
						10%	0.322	0.326	0.319	0.298	0.282	0.260	0.230
					CROSS_{III}^{ful}	*1%*	0.072	0.074	0.067	0.060	0.053	0.047	0.036
						5%	0.211	0.215	0.203	0.186	0.174	0.156	0.133
						10%	0.323	0.329	0.319	0.298	0.282	0.260	0.231

Continued on next page

Table IA.6 – *Continued from previous page*

	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	18.0	n.a.	168.0	n.a.	IND_I^c	*1%*	0.067	0.115	0.170	0.223	0.253	0.289	0.321
avg. t-stat	1.02	0.37	−0.04	0.28		*5%*	0.188	0.242	0.299	0.344	0.365	0.396	0.413
avg. α(%)	3.04	1.23	−0.05	0.83		*10%*	0.283	0.330	0.380	0.415	0.434	0.456	0.469
max t-stat	2.96	0.73	2.74	0.67									
max α *(%)*	12.81	7.49	14.78	7.95									
					IND_{II}^c	*1%*	0.069	0.117	0.173	0.224	0.253	0.290	0.321
						5%	0.191	0.244	0.299	0.344	0.366	0.396	0.413
						10%	0.285	0.332	0.380	0.415	0.434	0.456	0.468
					CROSS_I^c	*1%*	0.017	0.017	0.018	0.018	0.019	0.018	0.017
						5%	0.088	0.089	0.088	0.087	0.087	0.084	0.083
						10%	0.177	0.178	0.177	0.172	0.168	0.167	0.165
					CROSS_{II}^c	*1%*	0.073	0.078	0.068	0.059	0.053	0.047	0.036
						5%	0.218	0.224	0.208	0.187	0.173	0.158	0.135
						10%	0.337	0.343	0.328	0.304	0.286	0.264	0.234
					CROSS_{III}^c	*1%*	0.075	0.079	0.070	0.059	0.054	0.046	0.035
						5%	0.221	0.228	0.210	0.189	0.173	0.157	0.134
						10%	0.339	0.347	0.331	0.305	0.287	0.264	0.233

Table IA.7: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 0.5$, and fraction of outperforming funds $p = 5\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 0.5 and the fraction of outperforming funds of 5%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 0.5 to $p = 5\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.7 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	9.0	n.a.	177.0	n.a.	IND_{I}^{mis}	*1%*	0.057	0.077	0.112	0.149	0.172	0.204	0.230
avg. t-stat	0.93	0.45	−0.05	0.26		*5%*	0.165	0.194	0.223	0.252	0.271	0.295	0.314
avg. α(%)	2.96	1.84	−0.10	0.86		*10%*	0.253	0.279	0.299	0.325	0.335	0.350	0.367
max t-stat	2.55	0.78	3.02	0.95									
max α *(%)*	10.66	7.33	21.26	12.64									
					IND_{II}^{mis}	*1%*	0.059	0.077	0.113	0.150	0.173	0.203	0.230
						5%	0.166	0.195	0.224	0.252	0.272	0.295	0.315
						10%	0.254	0.281	0.300	0.325	0.336	0.351	0.367
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.004	0.010	0.010	0.010	0.009
						5%	0.011	0.011	0.034	0.047	0.051	0.053	0.052
						10%	0.039	0.041	0.082	0.101	0.106	0.109	0.111
					CROSS_{II}^{mis}	*1%*	0.061	0.062	0.050	0.041	0.037	0.033	0.022
						5%	0.180	0.184	0.160	0.138	0.129	0.115	0.095
						10%	0.283	0.285	0.261	0.230	0.217	0.198	0.172
					CROSS_{III}^{mis}	*1%*	0.061	0.064	0.050	0.042	0.037	0.033	0.023
						5%	0.182	0.185	0.161	0.139	0.127	0.114	0.095
						10%	0.285	0.288	0.264	0.231	0.217	0.198	0.171

Continued on next page

Table IA.7 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	6.7	1.3	132.7	5.8	IND^{ful}_{I}	*1%*	0.047	0.059	0.091	0.134	0.155	0.181	0.214
avg. t-stat	1.01	0.51	−0.04	0.28		*5%*	0.132	0.149	0.186	0.226	0.250	0.273	0.298
avg. α(%)	3.00	1.82	−0.07	0.85		*10%*	0.208	0.222	0.257	0.294	0.313	0.330	0.351
max t-stat	2.40	0.79	2.65	0.67									
max α *(%)*	8.63	5.94	13.62	7.55									
					IND^{ful}_{II}	*1%*	0.048	0.060	0.092	0.132	0.156	0.181	0.213
						5%	0.134	0.151	0.188	0.226	0.251	0.274	0.297
						10%	0.209	0.224	0.259	0.295	0.313	0.330	0.351
					CROSS^{ful}_{I}	*1%*	0.012	0.012	0.013	0.013	0.013	0.014	0.013
						5%	0.064	0.062	0.062	0.062	0.063	0.063	0.060
						10%	0.131	0.129	0.128	0.125	0.124	0.123	0.122
					CROSS^{ful}_{II}	*1%*	0.051	0.051	0.046	0.041	0.038	0.033	0.025
						5%	0.154	0.155	0.145	0.133	0.123	0.113	0.096
						10%	0.249	0.250	0.236	0.217	0.209	0.192	0.172
					CROSS^{ful}_{III}	*1%*	0.053	0.053	0.047	0.042	0.038	0.034	0.025
						5%	0.155	0.156	0.144	0.133	0.124	0.113	0.096
						10%	0.251	0.252	0.237	0.218	0.208	0.193	0.172

Continued on next page

Table IA.7 – *Continued from previous page*

	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%*	0.048	0.080	0.119	0.159	0.185	0.215	0.243
avg. t-stat	1.01	0.46	−0.04	0.28		*5%*	0.136	0.176	0.219	0.254	0.277	0.300	0.323
avg. α(%)	2.99	1.62	−0.07	0.83		*10%*	0.213	0.246	0.281	0.315	0.331	0.352	0.372
max t-stat	2.57	0.76	2.75	0.67									
max α *(%)*	9.66	6.27	14.86	7.90									
					IND_{II}^c	*1%*	0.050	0.082	0.121	0.159	0.185	0.216	0.244
						5%	0.138	0.177	0.218	0.255	0.276	0.300	0.324
						10%	0.215	0.247	0.283	0.316	0.331	0.352	0.370
					CROSS_I^c	*1%*	0.011	0.011	0.012	0.013	0.014	0.014	0.013
						5%	0.063	0.062	0.062	0.061	0.062	0.063	0.060
						10%	0.129	0.128	0.126	0.124	0.124	0.124	0.122
					CROSS_{II}^c	*1%*	0.054	0.053	0.045	0.040	0.037	0.035	0.025
						5%	0.159	0.163	0.147	0.132	0.124	0.115	0.096
						10%	0.258	0.258	0.240	0.221	0.210	0.195	0.170
					CROSS_{III}^c	*1%*	0.055	0.054	0.046	0.040	0.038	0.035	0.025
						5%	0.161	0.162	0.149	0.133	0.125	0.115	0.095
						10%	0.260	0.260	0.242	0.221	0.212	0.197	0.171

Table IA.8: **Simulated Test Power for $T = 60$ (1984–88), information ratio $IR = 0.5$, and fraction of outperforming funds $p = 2.5\%$**

Simulated test power for $T = 60$ (1984–88) and assumed information ratio of 0.5 and the fraction of outperforming funds of 2.5%. For all funds between 1984–88 that have at least eight observations, we collect their factor-adjusted 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 0.5 to $p = 2.5\%$ 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 as $\mathcal{D}^{mis}_{m,n}$) has 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 $\mathcal{D}^{ful}_m$. We different bootstrap methods to $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at a significance level (Sig. level) of 1%, 5%, and 10%. We study five bootstrap methods: IND_I is KTWW's baseline bootstrap where return residuals are resampled within each fund and factor realizations are kept intact; IND_{II} is KTWW's extended bootstrap where returns residuals are resampled within each fund and factor returns are sampled independently; $CROSS_I$ is FF's cross-sectional bootstrap where 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 $m = 1, 2, \ldots, 1000$ 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 alphas ('True') and zero alphas ('False')) such as the number of funds, the average (maximum) t-statistic, and the average (maximum) alpha. We calculate the mean and the standard deviation for these summary statistics across the simulation runs.

Continued on next page

Table IA.8 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	4.0	n.a.	182.0	n.a.	IND_{I}^{mis}	*1%*	0.051	0.067	0.091	0.126	0.148	0.178	0.202
avg. t-stat	0.94	0.62	−0.04	0.26		*5%*	0.143	0.165	0.186	0.220	0.239	0.262	0.283
avg. α(%)	3.05	2.65	−0.07	0.86		*10%*	0.222	0.240	0.258	0.280	0.297	0.314	0.331
max t-stat	2.05	0.85	3.05	0.96									
max α *(%)*	7.82	6.34	21.70	13.32									
					IND_{II}^{mis}	*1%*	0.051	0.066	0.093	0.127	0.148	0.178	0.202
						5%	0.143	0.167	0.187	0.220	0.239	0.261	0.282
						10%	0.224	0.242	0.259	0.279	0.297	0.314	0.330
					CROSS_{I}^{mis}	*1%*	0.001	0.001	0.004	0.007	0.009	0.009	0.009
						5%	0.011	0.010	0.027	0.038	0.043	0.044	0.046
						10%	0.036	0.035	0.067	0.085	0.091	0.094	0.095
					CROSS_{II}^{mis}	*1%*	0.054	0.054	0.041	0.033	0.030	0.026	0.020
						5%	0.158	0.155	0.132	0.116	0.107	0.098	0.080
						10%	0.249	0.247	0.221	0.202	0.187	0.172	0.149
					CROSS_{III}^{mis}	*1%*	0.054	0.055	0.041	0.034	0.031	0.026	0.020
						5%	0.158	0.157	0.132	0.117	0.108	0.098	0.081
						10%	0.252	0.249	0.224	0.203	0.187	0.173	0.149

Continued on next page

Table IA.8 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	3.0	0.8	136.4	5.8	IND^{ful}_{I}	*1%*	0.034	0.044	0.072	0.110	0.134	0.160	0.186
avg. t-stat	1.02	0.71	−0.04	0.29		*5%*	0.108	0.121	0.156	0.193	0.216	0.242	0.267
avg. α(%)	3.07	2.70	−0.05	0.84		*10%*	0.172	0.186	0.218	0.249	0.272	0.293	0.314
max t-stat	1.86	0.92	2.66	0.67									
max α *(%)*	6.17	5.12	13.87	7.80									
					IND^{ful}_{II}	*1%*	0.034	0.044	0.073	0.110	0.134	0.160	0.186
						5%	0.109	0.123	0.156	0.193	0.217	0.242	0.267
						10%	0.174	0.187	0.219	0.251	0.272	0.293	0.314
					CROSS^{ful}_{I}	*1%*	0.009	0.009	0.010	0.011	0.011	0.012	0.010
						5%	0.049	0.048	0.050	0.051	0.052	0.053	0.050
						10%	0.107	0.107	0.106	0.106	0.107	0.106	0.105
					CROSS^{ful}_{II}	*1%*	0.036	0.037	0.034	0.031	0.030	0.027	0.021
						5%	0.127	0.127	0.118	0.108	0.104	0.095	0.081
						10%	0.209	0.210	0.197	0.185	0.178	0.169	0.147
					CROSS^{ful}_{III}	*1%*	0.036	0.037	0.035	0.032	0.030	0.027	0.020
						5%	0.126	0.127	0.119	0.110	0.105	0.096	0.082
						10%	0.209	0.213	0.199	0.185	0.180	0.169	0.147

Continued on next page

Table IA.8 – *Continued from previous page*

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	4.0	n.a.	182.0	n.a.	IND_I^c	*1%*	0.036	0.060	0.096	0.135	0.157	0.188	0.217
avg. t-stat	1.01	0.61	−0.04	0.28		*5%*	0.112	0.142	0.181	0.219	0.239	0.266	0.292
avg. α(%)	3.07	2.25	−0.05	0.83		*10%*	0.177	0.205	0.239	0.272	0.292	0.314	0.333
max t-stat	2.10	0.83	2.76	0.67									
max α *(%)*	7.22	5.40	15.07	8.10									
					IND_{II}^c	*1%*	0.037	0.061	0.097	0.135	0.156	0.188	0.216
						5%	0.112	0.144	0.182	0.218	0.239	0.265	0.292
						10%	0.179	0.207	0.239	0.271	0.291	0.314	0.333
					CROSS_I^c	*1%*	0.009	0.009	0.010	0.011	0.011	0.012	0.010
						5%	0.049	0.048	0.048	0.050	0.054	0.054	0.051
						10%	0.106	0.103	0.106	0.106	0.106	0.107	0.105
					CROSS_{II}^c	*1%*	0.039	0.039	0.035	0.033	0.031	0.027	0.020
						5%	0.133	0.131	0.120	0.112	0.105	0.098	0.081
						10%	0.216	0.216	0.203	0.188	0.180	0.170	0.148
					CROSS_{III}^c	*1%*	0.040	0.040	0.036	0.033	0.031	0.027	0.020
						5%	0.134	0.133	0.122	0.112	0.106	0.097	0.081
						10%	0.220	0.219	0.204	0.188	0.181	0.170	0.149

B Five-Year Subsample, 2014–2018

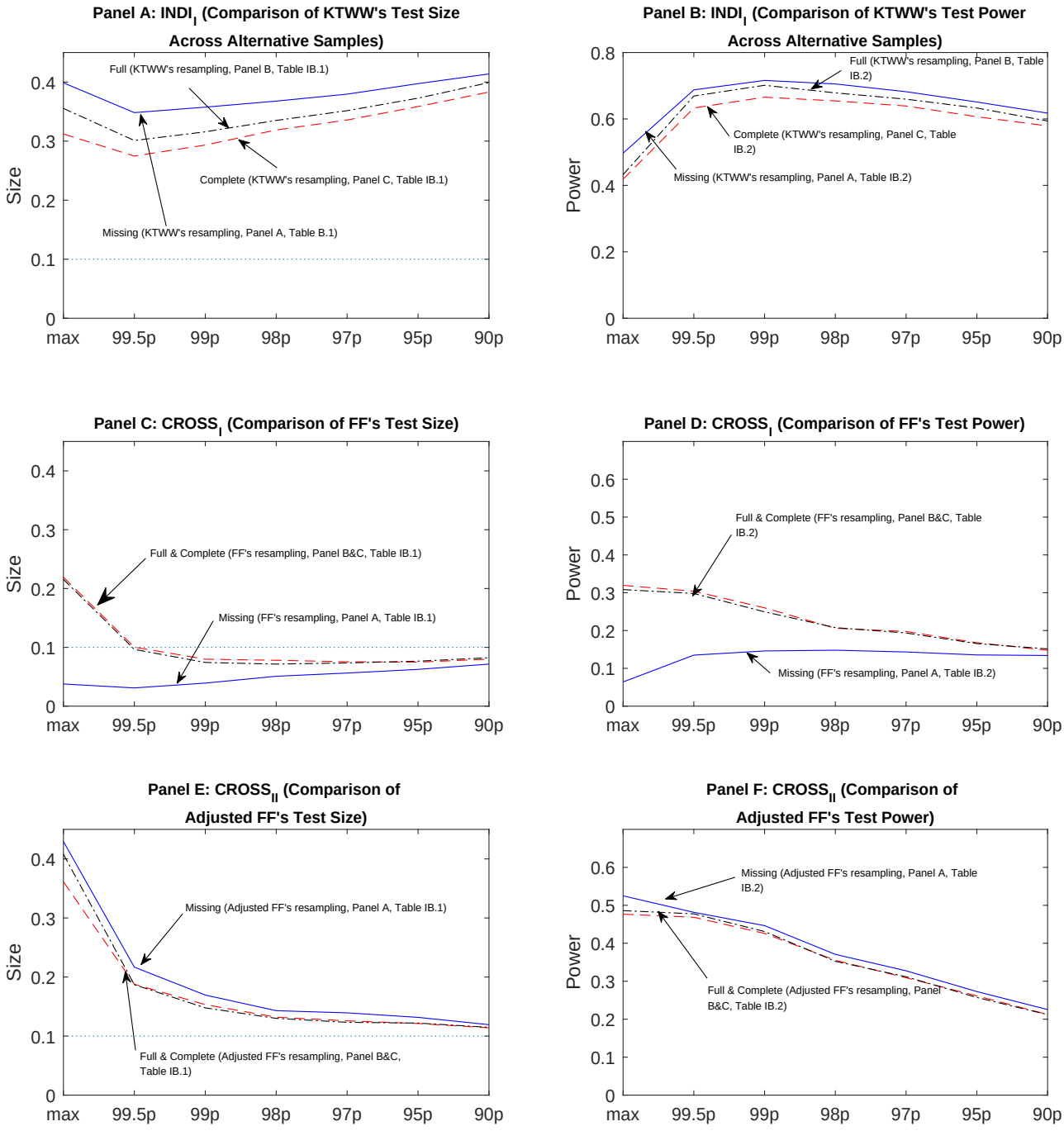


Figure IB.1: **Results: Test Size and Test Power, 2014–2018 (1,502 funds).** We report test size and test power at the 10% significance level. Test power corresponds to our baseline specification: $IR = 0.75$ and $p = 5\%$.

Table IB.1: **Simulated Test Size for $T = 60$ (2014–2018)**

For all funds between 2014–2018 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with a complete history of returns be $\mathcal{D}^{sub}$. We first inject an information ratio of $IR = 0$ into a fraction of $p = 0$ 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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.

Continued on next page

Table IB.1 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	n.a.	n.a.	1502.0	n.a.	IND_{I}^{mis}	*1%*	0.144	0.213	0.241	0.280	0.300	0.327	0.359
avg. t-stat	n.a.	n.a.	0.03	0.23		*5%*	0.309	0.296	0.318	0.337	0.354	0.373	0.394
avg. α(%)	n.a.	n.a.	0.02	0.38		*10%*	0.399	0.348	0.358	0.368	0.380	0.397	0.414
max t-stat	n.a.	n.a.	6.30	8.53									
max α *(%)*	n.a.	n.a.	24.80	16.47									
					IND_{II}^{mis}	*1%*	0.149	0.220	0.246	0.284	0.305	0.328	0.360
						5%	0.320	0.307	0.320	0.339	0.357	0.375	0.395
						10%	0.413	0.356	0.366	0.370	0.382	0.400	0.413
					CROSS_{I}^{mis}	*1%*	0.003	0.001	0.002	0.002	0.004	0.005	0.006
						5%	0.020	0.009	0.011	0.019	0.024	0.026	0.029
						10%	0.038	0.031	0.039	0.051	0.056	0.063	0.072
					CROSS_{II}^{mis}	*1%*	0.139	0.028	0.016	0.013	0.014	0.013	0.012
						5%	0.315	0.123	0.088	0.076	0.069	0.063	0.059
						10%	0.430	0.217	0.170	0.143	0.139	0.132	0.119
					CROSS_{III}^{mis}	*1%*	0.141	0.028	0.017	0.012	0.015	0.013	0.012
						5%	0.329	0.128	0.092	0.079	0.072	0.066	0.060
						10%	0.444	0.224	0.174	0.149	0.145	0.136	0.121

Continued on next page

Table IB.1 – *Continued from previous page*

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 B: Sample is $\mathcal{D}_{m,n}^{ful}$ (only funds with a complete history of returns)													
# of funds	n.a.	n.a.	1009.7	18.1	IND_I^{ful}	*1%*	0.141	0.165	0.192	0.233	0.255	0.283	0.324
avg. t-stat	n.a.	n.a.	0.03	0.24		*5%*	0.256	0.233	0.254	0.287	0.309	0.332	0.362
avg. α(%)	n.a.	n.a.	0.01	0.38		*10%*	0.312	0.275	0.293	0.319	0.336	0.359	0.383
max t-stat	n.a.	n.a.	4.06	1.19									
max α *(%)*	n.a.	n.a.	13.65	10.42									
					IND_{II}^{ful}	*1%*	0.148	0.166	0.196	0.234	0.259	0.285	0.324
						5%	0.267	0.238	0.260	0.291	0.312	0.335	0.362
						10%	0.318	0.281	0.298	0.323	0.337	0.360	0.386
					CROSS_I^{ful}	*1%*	0.046	0.005	0.005	0.005	0.005	0.007	0.007
						5%	0.143	0.039	0.026	0.029	0.030	0.034	0.033
						10%	0.219	0.100	0.080	0.078	0.076	0.075	0.080
					CROSS_{II}^{ful}	*1%*	0.153	0.025	0.015	0.012	0.012	0.014	0.011
						5%	0.287	0.111	0.080	0.068	0.064	0.060	0.053
						10%	0.361	0.188	0.153	0.132	0.126	0.121	0.114
					CROSS_{III}^{ful}	*1%*	0.163	0.026	0.014	0.014	0.013	0.014	0.011
						5%	0.297	0.113	0.080	0.071	0.067	0.060	0.054
						10%	0.367	0.194	0.157	0.133	0.128	0.124	0.116

Continued on next page

Table IB.1 – *Continued from previous page*

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	n.a.	n.a.	1502.0	n.a.	IND_I^c	*1%*	0.152	0.195	0.220	0.260	0.285	0.310	0.348
avg. t-stat	n.a.	n.a.	0.03	0.24		*5%*	0.299	0.263	0.284	0.308	0.329	0.353	0.379
avg. α(%)	n.a.	n.a.	0.01	0.37		*10%*	0.355	0.301	0.316	0.335	0.352	0.373	0.399
max t-stat	n.a.	n.a.	4.35	1.25									
max α *(%)*	n.a.	n.a.	15.94	11.83									
					IND_{II}^c	*1%*	0.160	0.200	0.224	0.265	0.285	0.312	0.349
						5%	0.309	0.270	0.289	0.311	0.331	0.355	0.378
						10%	0.367	0.307	0.318	0.337	0.354	0.374	0.400
					CROSS_I^c	*1%*	0.041	0.004	0.004	0.004	0.006	0.006	0.006
						5%	0.139	0.035	0.028	0.031	0.034	0.032	0.031
						10%	0.216	0.097	0.074	0.072	0.074	0.077	0.082
					CROSS_{II}^c	*1%*	0.163	0.023	0.014	0.011	0.013	0.013	0.011
						5%	0.330	0.109	0.076	0.066	0.064	0.059	0.052
						10%	0.408	0.187	0.148	0.130	0.123	0.122	0.115
					CROSS_{III}^c	*1%*	0.174	0.024	0.014	0.013	0.013	0.014	0.010
						5%	0.339	0.110	0.079	0.068	0.064	0.060	0.055
						10%	0.418	0.195	0.153	0.133	0.129	0.124	0.115

Table IB.2: **Simulated Test Power for $T = 60$ (2014–18), information ratio $IR = 0.75$, and fraction of outperforming funds $p = 5\%$**

For all funds between 2014–2018 that have at least eight observations, we collect their returns into a data matrix $\mathcal{D}$. Let the corresponding return matrix for funds with 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 as $\mathcal{D}^{mis}_{m,n}$) 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 $\mathcal{D}^{ful}_m$. We use different methods to bootstrap $\mathcal{D}^{mis}_{m,n}$, $\mathcal{D}^{ful}_m$, and $\mathcal{D}^c_m$, respectively, at significance levels (Sig. level) 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 alphas (True) and zero alphas (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.

Continued on next page

Table IB.2 – *Continued from previous page*

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 $\mathcal{D}_{m,n}^{mis}$ (including missing observations)													
# of funds	75.0	n.a.	1427.0	n.a.	IND_{I}^{mis}	*1%*	0.062	0.446	0.541	0.565	0.562	0.552	0.536
avg. t-stat	1.56	0.28	0.03	0.26		*5%*	0.211	0.582	0.632	0.644	0.629	0.606	0.586
avg. α(%)	2.85	0.50	0.02	0.39		*10%*	0.331	0.650	0.688	0.685	0.668	0.639	0.612
max t-stat	5.17	2.19	6.24	9.29									
max α *(%)*	14.26	7.34	25.35	18.75									
					IND_{II}^{mis}	*1%*	0.052	0.458	0.545	0.573	0.567	0.556	0.541
						5%	0.204	0.597	0.644	0.649	0.633	0.608	0.586
						10%	0.333	0.664	0.698	0.692	0.671	0.641	0.612
					CROSS_{I}^{mis}	*1%*	0.000	0.000	0.003	0.007	0.008	0.008	0.010
						5%	0.000	0.007	0.029	0.044	0.050	0.053	0.060
						10%	0.000	0.026	0.087	0.114	0.120	0.124	0.127
					CROSS_{II}^{mis}	*1%*	0.081	0.124	0.074	0.047	0.037	0.031	0.026
						5%	0.245	0.321	0.261	0.207	0.174	0.142	0.114
						10%	0.383	0.462	0.432	0.363	0.323	0.270	0.224
					CROSS_{III}^{mis}	*1%*	0.074	0.130	0.076	0.051	0.042	0.034	0.027
						5%	0.247	0.330	0.270	0.212	0.179	0.145	0.117
						10%	0.387	0.481	0.451	0.372	0.329	0.275	0.229

Continued on next page

Table IB.2 – *Continued from previous page*

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 B: Sample is $\mathcal{D}^{ful}_{m,n}$ (only funds with a complete history of returns)													
# of funds	50.4	4.1	958.8	17.7	IND^{ful}_{I}	*1%*	0.250	0.442	0.516	0.535	0.523	0.514	0.502
avg. t-stat	1.71	0.32	0.03	0.25		*5%*	0.361	0.567	0.611	0.612	0.595	0.578	0.552
avg. α(%)	2.84	0.50	0.01	0.39		*10%*	0.418	0.634	0.665	0.653	0.638	0.606	0.579
max t-stat	4.75	1.94	4.04	1.18									
max α *(%)*	11.18	5.72	13.59	10.40									
					IND^{ful}_{II}	*1%*	0.262	0.455	0.524	0.537	0.527	0.517	0.506
						5%	0.374	0.578	0.621	0.619	0.602	0.580	0.552
						10%	0.428	0.645	0.671	0.656	0.641	0.609	0.580
					CROSS^{ful}_{I}	*1%*	0.123	0.037	0.020	0.014	0.015	0.013	0.013
						5%	0.242	0.173	0.129	0.098	0.086	0.074	0.072
						10%	0.321	0.308	0.262	0.208	0.198	0.168	0.149
					CROSS^{ful}_{II}	*1%*	0.263	0.133	0.080	0.046	0.039	0.030	0.023
						5%	0.395	0.332	0.263	0.196	0.173	0.143	0.113
						10%	0.478	0.471	0.427	0.357	0.311	0.261	0.213
					CROSS^{ful}_{III}	*1%*	0.274	0.138	0.084	0.051	0.039	0.032	0.022
						5%	0.404	0.337	0.275	0.204	0.177	0.144	0.112
						10%	0.492	0.485	0.441	0.363	0.316	0.261	0.217

Continued on next page

Table IB.2 – *Continued from previous page*

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	75.0	n.a.	1427.0	n.a.	IND_I^c	*1%*	0.266	0.517	0.573	0.585	0.571	0.551	0.535
avg. t-stat	1.71	0.30	0.03	0.25		*5%*	0.383	0.621	0.657	0.650	0.628	0.606	0.574
avg. α(%)	2.84	0.47	0.01	0.39		*10%*	0.432	0.669	0.701	0.676	0.657	0.631	0.594
max t-stat	5.11	2.13	4.33	1.23									
max α *(%)*	12.54	6.80	15.92	11.97									
					IND_{II}^c	*1%*	0.275	0.528	0.581	0.591	0.576	0.554	0.536
						5%	0.393	0.629	0.666	0.652	0.632	0.607	0.575
						10%	0.444	0.684	0.707	0.680	0.660	0.633	0.596
					CROSS_I^c	*1%*	0.121	0.034	0.022	0.016	0.015	0.014	0.015
						5%	0.228	0.172	0.127	0.096	0.086	0.074	0.072
						10%	0.309	0.300	0.252	0.210	0.193	0.166	0.152
					CROSS_{II}^c	*1%*	0.273	0.134	0.080	0.048	0.040	0.034	0.028
						5%	0.410	0.33	0.258	0.198	0.171	0.136	0.108
						10%	0.486	0.480	0.430	0.353	0.313	0.258	0.213
					CROSS_{III}^c	*1%*	0.287	0.139	0.084	0.049	0.039	0.032	0.025
						5%	0.423	0.345	0.268	0.205	0.179	0.142	0.111
						10%	0.499	0.494	0.438	0.360	0.319	0.261	0.216