Testing Asset Pricing Models on Individual Stocks

Charles Clarke[*] Morteza Momeni[†]

January 2023

Abstract

This paper tests asset pricing models using individual stocks as test assets, rather than sorted portfolios. Sorted portfolios have the severe limitation that the researcher must know, in advance, reliable predictors of expected returns. We show how to generate appropriately sized tests and verify that our tests have considerable test power. In simulations when the CAPM describes the population, our tests (correctly) reject the Fama and French (2015) six factor model 97.5% of the time, while our tests (incorrectly) reject the CAPM less than 5%. We apply our tests to several leading factor models and reject nine of the eleven models tested. The instrumented factor model of Kelly et al. (2019) stands out as the most successful.

JEL classification: G12, C15

[*] University of Kentucky, Gatton School of Business, Department of Finance and Quantitative Methods, Lexington, KY 40506; charlie.clarke@uky.edu

[†] University of Kentucky, Gatton School of Business, Department of Finance and Quantitative Methods, Lexington, KY 40506; morteza.momeni@uky.edu

1 Introduction

Empirical asset pricing models are primarily tested on portfolios, rather than on individual stocks. This requires the researcher to take a stand on which sorting variables are reliable predictors of expected returns across stocks. The rising concern that many of these characteristics are unreliable out-of-sample casts doubt on these tests. If the characteristics are the result of data snooping, then a true model may be rejected.

We test asset pricing models using individual stocks. We compare the cross-sectional alpha across individual stocks to thousands of new bootstrapped histories generated from an adjusted version of the original sample, altered so that the alpha is zero. The new histories capture the statistical structure of the original data but generate new distributions of the cross-sectional alpha that would be observed under the null hypothesis.

This procedure is an adaptation of methodology developed by Kosowski et al. (2006) and Fama and French (2010) in the fund performance literature. Their insight is that we can evaluate whether some fund managers have “skill” without having to pre-specify the predictors of that skill. Skill is alpha relative to a model in the mutual fund literature just as mispricing is alpha relative to a model in the asset pricing literature. Just as managers will outperform or underperform their benchmarks, stocks will outperform or underperform an asset pricing model. But by comparing sample alpha generated by this random variation to data generated by a simulated population, otherwise identical, where all alphas are known to be zero, we can evaluate whether the sample alpha is sufficient to reject that the true alpha is zero.

This approach allows us to test asset pricing models without taking a stand on the hundreds of stock pricing characteristics (Harvey et al., 2016) and whether they were data snooped (Lo and MacKinlay, 1990). Whereas, in traditional tests, if the models are tested on characteristic sorted portfolios that are formed on ex post performance, rather than ex ante differences in expected returns, the true asset pricing model will likely be erroneously rejected. Our methods do not require researchers to make the difficult choice about which characteristics will generate reliable test portfolios.

Individual stocks raise several issues. Individual stocks have non-normal distributions and dependent correlation structures that can make generating tests of the appropriate test size difficult. We show these concerns are justified. We simulate separate populations, where the CAPM or Fama French six factor (FF6) model perfectly describe the population. When the Fama and French (2010) procedure is adapted without alteration, the resulting test sizes for a traditional 5% test are 23% for the CAPM and 26% for the FF6 model. The procedure rejects the true model 23% to 26% of the time. We show how to expand the confidence intervals to find tests of appropriate size.

Additionally, individual stocks have poorly estimated betas that raise issues of statistical power. After constructing confidence intervals of the appropriate size, we show that our tests remain powerful. In our simulations, even when our most conservative adjustment is applied guaranteeing test size below 5%, the procedure correctly rejects the FF6 model when the CAPM is the true model 97.5% of the time. When the population follows the FF6 model, the procedure correctly rejects the CAPM 83.2% of the time. The simulations suggest these corrections are excessively conservative leading to a test size well below 5%. We also explore optimized confidence intervals that maximize power, while maintaining a given test size. These optimized confidence intervals increase test power to 99.9% and 91.0%, respectively.

We proceed to test several of the leading asset pricing models in the field. These models include the CAPM (Sharpe, 1964), the Fama and French factor models, including the three factor model with size and value factors (Fama and French, 1993), the five factor model that adds profitability and investment factors (Fama and French, 2015), a six factor model that adds the momentum factor of Carhart (1997), and the five factor model of Pástor and Stambaugh (2003) that combines size, value and momentum factors with a traded liquidity factor. Additionally, we test two versions of the Q-factor model, a four factor version of Hou et al. (2015) with market, size, investment and profitability factors and a five factor version that adds an expected investment growth factor (Hou et al., 2021). We also test the four factor model of Stambaugh and Yuan (2017), which combines the market and size factors with two factors, mgmt (management) and perf (performance), formed to capture mispricing. Lastly, we test three versions instrumented principal components (IPCA) factors of Kelly et al. (2019). Each model is six factors. The restricted and unrestricted versions differ in that the former creates IPCA factors restricts the explanatory power of the characteristics to be through the common factors. The models are similar in that they are both formed using the entire sample. The third version, out-of-sample, forms IPCA factors using an expanding window limiting the concern for look-ahead-bias in the model parameters.

The star of these tests is the IPCA model of Kelly et al. (2019). When tested on individual stocks, all other models are rejected, even with our most conservative thresholds. The IPCA restricted model, and importantly, the out-of-sample IPCA model, both pass our tests, even at our less conservative thresholds, where the statistical power to reject is highest.

We develop a simple measure of combined mispricing, which captures the difference from the observed t-statistics of alpha at each decile from the alpha expected under the null hypothesis. We reject when this distance is so large it is out of the confidence interval. Our composite mispricing measure captures the combined differences from the realizations expected under the null hypothesis of zero alphas.

Next, we revisit our tests removing the small, but plentiful micro-cap stocks from our sample. Several models perform better on this subsample, especially the Fama and French style models. We still reject the CAPM and, interestingly, a six factor Fama and French model that adds momentum to the five factor version, but we do not reject the three or five factor models in this sample. That we can add factors to a model and observe its performance deteriorate demonstrates the benefit of these tests relative to others explored in the literature. Barillas and Shanken (2017, 2018) show that many tests, when factors are appropriately included as test assets, devolve into tests of which set of factors generate larger ex post Sharpe ratios. But in the presence of data snooped characteristics, this conclusion may lead to a problematic cycle of adding high ex post Sharpe ratio factors that were not ex ante predictors of expected return. The true asset pricing model would be rejected for not including these snooped factors and would not be able to price managed portfolios formed on these snooped characteristics.

Individual stocks provide a check on this cycle. False factors formed on ex post alphas may perform well at pricing portfolios formed on the same or similar characteristics that capture these ex post alphas (Ferson et al., 1999, 2003), but these factors should not perform well at pricing individual stocks. We demonstrate an extreme version of this by creating an extremely high ex post Sharpe ratio factor model from a set of hundreds that Hou et al. (2015) explored. Given that the model appeals to no theory and is chosen only on ex post performance measures, we think it highly likely to be data snooped. We show, despite its high Sharpe ratio, our model can be rejected due to its substantial mispricing of individual stocks.

Our paper is indebted to a considerable and growing literature exploring fund performance. Kosowski et al. (2006) and Fama and French (2010) pioneer this bootstrap approach in the fund performance literature. The central question in these papers, do fund managers have skill? has a clear analogue to our question, does a factor model price stocks? as skill and mispricing are both defined as alpha relative to a model. Despite this close connection, the efficacy of this methodology has not been evaluated in the context of asset pricing tests across individual stocks. We show when properly applied this methodology generates appropriate size and considerable power in asset pricing tests of factor models on individual stocks. These tests are close in spirit to the size and power tests Harvey and Liu (2020) use to reevaluate and reconcile early contradictions of Kosowski et al. (2006) and Fama and French (2010), the latter finding no evidence for fund manager skill, in contrast to the former. The fund performance literature has expanded to identifying the distribution of skill (Barras et al., 2010; Chen et al., 2017; Ferson et al., 2019) as well as detecting skill out-of-sample (Harvey and Liu, 2018; Giglio et al., 2019).

Harvey and Liu (2019) make an important first attempt to bring this literature into the cross-section of stock returns. These authors ask, starting with a set of candidate factors and adding each factor one at a time, when is the growing factor model no longer significantly better at explaining alpha after controlling for multiple testing? When should we stop adding an additional factor from a larger set? This question aligns closely with their goal of narrowing down a large set of candidate factors, motivated by the growing characteristics "zoo" documented by Harvey et al. (2016). A drawback of this approach is that, due to the adjustments for multiple hypotheses, the threshold for retaining a factor is dependent on how many factors one starts with. Adding irrelevant factors raises the threshold for an additional factor to significantly improve the overall model and can result in a smaller number of retained factors.

Our question stems from the literature on testing asset pricing models. Does a set of factors explain a set of test assets or are the pricing errors sufficiently large that we can reject the model? In this setting, the models are self-contained and rely on different theoretical motivations. Whereas, it may not be theoretically coherent to combine a mispricing factor of Pástor and Stambaugh (2003) with an investment factor of Hou et al. (2015) and an IPCA factor of Kelly et al. (2019). We show, contrary to the thrust of the literature, that this adapted methodology generates tests on individual stocks that have appropriate size and considerable power.

Also inspired by the fund performance literature, Barras (2019) adapts the false discoveries approach of Barras et al. (2010) to measure the proportion of mispricing on micro portfolios that contain only a small number of stocks. This approach captures some of the benefits of portfolios in estimating betas, as well as the broad spread betas generated by individual stocks. But the approach still requires pre-specifying characteristics to create the micro portfolios, and Barras (2019) shows the false discoveries approach has low power on individual stocks.

Considerable progress has been made taking cross-sectional asset pricing approaches to individual stocks. These papers pre-specify factors and explore whether the betas on these factors align with average returns, generating a risk premium. Estimation error in the betas creates an errors-in-variables problem that biases the betas downward (Blume, 1970). Ang et al. (2020) argues that while the measurement errors in the betas are larger across individual stocks, the larger cross-sectional spread in the betas more than offsets this error leading to more precise risk premium estimates in individual stocks. Jegadeesh et al. (2019) tackle the errors-in-variables problem directly using an instrumental variables approach that uses in sample beta estimates as instruments for out-of-sample betas. Raponi et al. (2020) explore large N, fixed T cross-sectional tests of beta-pricing models on individual stocks.

Gagliardini et al. (2016) extend the two-pass regression methodology to large panels with time-varying risk premiums and loadings. Chaieb et al. (2021) apply this methodology to asset pricing models in a large panel of international stocks. The two pass approach requires the tested model to specify all of the asset pricing factors, even if the factors are not priced. If omitted factors are present, the risk premia of the second pass estimates do not converge to the risk premia of the priced factors (Gagliardini et al., 2016; Gagliardinia et al., 2020). While Gagliardini et al. (2019) suggest a diagnostic test for the presence of omitted factors, this is a heavy burden to impose on model specification. The set of potential priced factors is already immense (Harvey et al., 2016), but the set of unpriced factors is may be even larger. This intuition is captured by Roll and Ross (1984), “firms are in industries together, or inhabit the same region of the country, or produce substitute or complement products, or compete for the same labor, etc...We expect there are as many factors as there are sets of assets.” Since our methodology utilizes the time-series tests, rather than the two-pass approach, our tests are valid in the presence of omitted factors. This alternative approach provides a complement to the growing literature on cross-sectional asset pricing tests in large panels.

While tests of characteristic portfolio sorts have valuable uses, they have the clear drawback of having to prespecify reliable predictors of expected returns. There is no consensus on which of the hundreds of return predictors are reliable. Additionally, important drivers of cross-sectional variation in expected returns may remain unknown. By dis-aggregating tests of factor models down to the level of individual stocks, our tests provide an alternative that does not require adjudicating these difficult questions.

2 Methodology

We start with a sample of N stocks captured by a statistical factor model (represented here as a one factor model for ease of explication).

$$R_{i,t} = \alpha_i + \beta_{i,1} F_{1,t} + \epsilon_{i,t}$$

We test the traditional null hypothesis for asset pricing that all of the (population) alphas equal zero for the N stocks.

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_N = 0$$

Even if the population alphas are zero and the model explains all of the variation in average returns across stocks, we know that the observed, in-sample alphas will not be zero in any finite sample. We test the null hypothesis by creating a pseudo-population, in which the null hypothesis is true, that is otherwise identical to the sample. By resampling from this pseudo-population, we simulate data generated under the null hypothesis. Then we can ask, are the sample observed alphas consistent with the null hypothesis being true? or are the observed alphas so significantly different than those simulated, that the null hypothesis is unlikely to be true?

We create an empirical distribution by first running N time-series regressions to estimate the in-sample alphas and betas of our N stocks. We create a new "population" by subtracting the estimated alphas from the sample of stock returns.

$$Z_{i,t} = R_{i,t} - \widehat{\alpha_i} = \widehat{\beta_{i,1}} F_{1,t} + \widehat{\epsilon_{i,t}}$$

This creates a panel of N stocks and T time periods. We resample from this panel by drawing one of the T time-periods from the panel and retaining the entire cross-section of stocks in that time period and the observed realizations of the candidate model's factors. We continue this cross-sectional bootstrap T times to create a new sample. This sample is consistent with the null hypothesis, but retains the cross-sectional dependencies in the original data. We repeat the time-series regressions on our new sample of stocks and generate new alpha estimates.

We condense this new set of alphas into a set of P percentiles. This allows us to ask, is the alpha we observe at a given percentile in the sample data consistent with the alpha we would expect to observe at that percentile if the null hypothesis was true? Since we test the null hypothesis at P separate percentiles, we adjust the individual tests for multiple hypotheses to achieve the appropriate aggregate test size. Our most conservative criterion applies the Bonferroni correction, taking the original test size (α_s) and dividing by P tests. The Bonferroni correction is conservative. It guarantees the aggregate adjusted test size is no greater than the acceptable level, but the resulting test size can be considerably less. An overly conservative correction can result in a decline in test power, the ability to reject a false null hypothesis. We also explore an alternative approach that we call optimized confidence intervals. In this approach, we progressively lengthen the confidence intervals uniformly across percentiles, until our desired test size of 5% is reached. These optimized confidence intervals maximize the power of our tests given a desired test size.

The tests have thus far been described in terms of sample and resampled alpha, but Fama and French (2010) focus their attention on the alphas' t-statistics. This approach has the advantage of controlling for residual risk, as well as accounting for variation in the sample size across stocks. Our simulations tackle both approaches, while our main empirical results focus attention on the t-statistics.

3 Simulations of Methodology

In this section, we describe simulations that verify our methodology and explore its statistical power. Rather than construct artificial stock data, we adjust real stock data to create the appropriate statistical setting, while preserving much of the structure of the original data.[1] We explore the size and power of the methodology in two different settings. In the first, we construct a population in which the data generating process is the CAPM. When the CAPM is the candidate model, the null hypothesis holds and the size of the test is how often the method rejects the true model. We also test a Fama and French six factor model on the same sample, that includes the market, size (smb), value (hml), profitability (rmw), investment (cma) and momentum (umd) factors. Since the population is described by the CAPM, when the FF6 model is the candidate, the null hypothesis is false and the power of the test is how often we reject this false null hypothesis. Next, we reverse the example creating a population that is described by the FF6 model. Now when the CAPM is the candidate model, the simulation explores the test's power to correctly reject a false model, and when the FF6 model is the candidate model, the simulation explores the resulting test size, incorrectly rejecting the true model.

3.1 Simulation Procedure

We construct the zero-alpha population by first estimating the alpha for all the stocks in our data and then subtracting this estimated alpha from the series of returns. For the CAPM, we regress excess returns on the market factor in a series of N time-series regressions.

$$R_{it} = \alpha_i + \beta_i mktrf_t + \epsilon_{it}$$

We subtract the estimated alpha from our sample to create a zero-alpha population (ZAP) of stock returns.

$$R_{it} - \widehat{\alpha}_i = ZAP_{it}$$

We then resample T periods with replacement from this zero-alpha population to create a simulated sample of stocks. We also retain the factor realizations for these T periods.

Now, with this sample generated under a known data generating process, we specify a candidate factor model, either the CAPM or FF6 model, and test the null hypothesis that the alphas of all stocks are jointly zero. We then proceed to estimate the sample observed

[1] Harvey and Liu (2020) use a similar simulation approach to compare methods in the fund performance literature.

alpha given the candidate factor model using the returns and factors from the simulated sample. We then subtract this estimated alpha to create a sample data set true under the null hypothesis. We resample cross-sectionally T periods with replacement from this pseudo-population to generate test statistics. From our zero-alpha population of stocks, we create 1000 samples. For each individual sample, we bootstrap 1000 new samples under the candidate null hypothesis. In total, 1,000,000 (1000 by 1000) samples under each of two data generating processes (CAPM and FF6) and two candidate models (CAPM and FF6).

3.2 Simulation Data

Our simulations are built from a large sample of U.S. stocks that are modified to create a new population, and then resampled to preserve the original samples important features, such as cross-sectional dependencies and extreme return realizations. Our sample consists of stock returns from the Center for Research in Security Prices (CRSP). Our sample starts in July of 1964 and ends in December of 2018. We restrict our sample to common equity (shrcd 10 or 11) from the major exchanges, New York Stock Exchange, NASDAQ and NYSE America (exchcd 1, 2 or 3). To be included in our sample a stock must have a share price and outstanding shares in the month before the return date and at least 36 non-missing returns.

Table 1 shows summary statistics of the data in our sample. The first four columns show the results of time-series regressions on individual stocks. We aggregate at the stock level and report nine percentiles, displayed in the first column. The second through fourth columns show R-squareds, alphas and t-statistics for both the CAPM in Panel A and the FF6 model in Panel B. These three columns capture several the difficulties in asset pricing tests tests across individual stocks. The spread in explained variation is large with 10% having R-squareds smaller than 1.1% and 10% have R-squareds larger than 27.4% for the CAPM. The FF6 model generates a smaller, but still large, 10$^{\text{th}}$ percentile to 90$^{\text{th}}$ percentile range of 7.9% to 40.1%. The low explained variation lead to alphas and t-statistics that are estimated with considerable error making traditional asset pricing tests infeasible or uninformative. The next column shows large spreads in sample alphas. The third column of t-statistics, show that these large alphas are measured with considerable error. The t-statistic range from the 10$^{\text{th}}$ percentile to 90$^{\text{th}}$ percentile of -1.61 to 1.57 is not obviously different than expected from chance alone.

The last two columns show the transformed data. We add or subtract a constant to each stock's monthly returns to create a population with zero alpha for a given model without changing the percent of variation explained. In Panel A, the sample has been transformed to have zero alpha under the CAPM, so the last two columns show the remaining alphas and t-statistics for the FF6 model. In Panel B, the sample has been transformed to have zero alpha under the FF6 model, so the last two columns show the remaining CAPM alpha and t-statistics. These last two columns represent new populations that will be used to test size and power. The t-statistics in the last column again demonstrate the challenge of evaluating asset pricing models on individual stock returns. The 10th to 90th percentile of t-statistics range from -0.85 to 1.07 for the CAPM and from -1.06 to 0.75 for the FF6 model. Our simulations will explore whether it is possible to sort through the considerable uncertainty in these estimates across individual stocks to generate appropriately sized tests that remain reasonably powerful.

3.3 Simulation Results

Our simulation data retains the cross-sectional dynamics and resulting challenges of testing asset pricing models on individual stocks. Nevertheless, we are able to proceed by aggregating information across the distribution of observed alphas and comparing to the distribution of observed alphas expected under the null hypothesis. We draw (with replacement) 654 months of factors and stocks from our population of stocks transformed to be fully described by the CAPM (or FF6 model) to generate a single sample. We treat this as if it were the data actually observed from July 1964 to December 2018.

We take this data through our procedure. First, we estimate the alphas observed for each stock with at least 36 observations using time-series regressions on the candidate model. These are our sample alphas for this simulated history. Then, we generate a new pseudo-population from this history of returns by subtracting these sample alphas. This generates a new pseudo-population of stock returns fully described by the null hypothesis for each candidate model. We bootstrap from this pseudo-population to generate a new sample and factors and use this sample to estimate the observed alphas. We repeat this resampling procedure 1000 times. This simulates one run of a researcher receiving a history of returns and then testing a candidate model. We repeat by resampling a new history of returns until we reach 1000 simulations. With 1000 potential histories and 1000 bootstraps of each history, the simulation requires 1,000,000 total bootstrap samples.

We condense the sample history of data from July 1964 to December 2018 to 9 observed deciles. We condense each of the 1000 bootstrap samples generated under the null hypothesis of no alpha as well. By comparing the sample observed alpha at a given percentile to the percentiles generated by our 1000 simulations, we generate test statistics to evaluate whether the sample data is significantly different than that expected under the null hypothesis. If a given percentile is out of the range generated by our simulations, then we reject that the sample data is consistent with the null hypothesis. Since we are testing a hypothesis at several deciles, we must increase the 95% confidence interval in a way that preserves the 5% overall size of our test.

The left panel of Figure 1 shows the results of the very first simulated history. We have generated a history of returns from a population constructed so that the CAPM holds. In this population, CAPM alphas are zero for all stocks. The panel shows the results when the CAPM is the candidate model being tested. The blue dots show the t-statistics of alphas at a given decile of time-series regressions of the stock returns on the excess market return. The t-statistics of alphas are shown for the 10th percentile, 20th percentile, 30th percentile and so on up to the 90th return percentile. The gray lines are the confidence intervals generated under the null hypothesis that stocks have CAPM alpha. Since the hypothesis is being tested at nine points, the confidence intervals are adjusted using the Bonferroni correction for multiple hypotheses to 99.44%, wider than the 95% required for one test[2]. This guarantees an overall test size of 5% or less.

In the first simulation, the 40th percentile stock has a CAPM t-statistic of -0.18. This value is well within the 99.44% confidence interval the generated under the null hypothesis that the CAPM is true, which is -0.56 to -0.04. In this example, the CAPM null is not rejected at the 40th percentile. After we examine the other eight deciles, we see that it is not rejected at any point. The CAPM candidate model has failed to be rejected under the null hypothesis that the CAPM describes the data. Since, by construction, the CAPM describes all average returns in population, this is a success. An appropriately sized test would only reject the true model 5% of the time. In contrast, when that same history of

[2] The Bonferonni correction takes the total desired size, 5%, and divides by the number of hypothesis tests, nine, giving an adjusted size of $\frac{5\%}{9} = 0.555\%$. This yields an adjusted confidence interval of 99.44%.

returns is tested against the FF6 candidate model, as shown in the right panel of Figure 1, we reject the null hypothesis that the FF6 model is true. Since the candidate model is false, the CAPM describes the sample returns, not the FF6 model, this is a success. Specifying the null hypothesis with a false candidate model tests the power of our procedure. To test models, we require a procedure that fails to reject the true null hypothesis, but succeeds at rejecting false null hypotheses.

In our simulations, we create two populations. One in which the CAPM describes expected returns and one in which the FF6 model describes expected returns. Additionally, we specify the null hypothesis relative to both candidate models. This setup gives us two tests of size, one when the CAPM holds in population and the CAPM is the candidate model and one when the FF6 model holds in population and the FF6 is the candidate model, and two tests of power, when the CAPM describes the population, while the FF6 is the candidate model and conversely when the FF6 describes the population and the CAPM is the candidate model.

Table 2 shows the result of these four cases. The top panel shows the results, when the conservative Bonferroni corrections are used. The top of the panel shows the CAPM and FF6 candidate models tested against CAPM population at the 9 deciles between the 10^{th} and 90^{th} percentiles. Using the t-statistic as the test statistic of interest, the CAPM is rejected 1.5% of the time. We require a procedure to reject at least 5% of the time, but the Bonferroni corrections are typically conservative, which lowers the size of the test. The next row shows that the false candidate model, FF6, is rejected 97.5% of the time. Even with the conservative Bonferroni corrections, the procedure retains considerable test power. When the alpha statistic is used the size of the test is 3.2% and the power is 98.8%. Again, this is within the required range, though conservative for test size, but retains considerable test power.

The next panel shows the same size and power tests, when the FF6 model describes the population. Now, the FF6 model is the true candidate model and the CAPM is the false candidate model. When the FF6 model is the true model it is rejected 3.3% of the time when the t-statistic is the statistic used and 4.9% when the alpha is the test statistic used. The model rejects the false CAPM 83.2% of the time using the t-statistic and 78.8% of the time using the alpha.

The next panel of Table 2 shows the Optimized Confidence Interval. These confidence intervals maximize test power for a given test size. We gradually decrease the length of the confidence interval uniformly across the percentiles, until the test size is close to 5%. We search across a discrete grid, so the match is only approximate.[3] Comparing the top and bottom panels, we see that the optimized confidence intervals do increase test power. When the CAPM describes the population, the already high FF6 test power increases from 97.5% to 99.9% t-statistics are the test statistic and from 98.8% to 99.9% when alphas are the test statistic. When the FF6 model describes the population, the CAPM test power increases from 83.2% to 91% t-statistics are the test statistic and from 78.8% to 81.2% when alphas are the test statistic.

Figure 2 captures the trade-off in test size and test power across different confidence intervals. The left panel shows the results when the CAPM generates the population and the right panel shows the results when the FF6 model generates the population. The x-axis is the length of the confidence interval. The left y-axis is the size of the test (how often the true null hypothesis is rejected) and the right y-axis is the power of the test (how often the false model is rejected). In both figures, test size declines linearly as confidence intervals widen, while test power declines non-linearly. The dash-dotted line in the figure displays the Bonferroni correction mandates and shows it enforces a conservative test size, below 5% in both panels. We uniformly narrow the confidence intervals so that the test size is exactly 5%. This approach will yield greater test power. The dashed line in the figures shows this optimized confidence interval. In both figures, the optimized confidence interval yields additional test power.

Next, we expand the percentiles from 9 to 99. This yields a finer grid over which to search for mispricing and extends the search further into the tails. It is infeasible to attempt a Bonferroni adjustment as the confidence intervals generated will be extremely wide with this number of test assets. The Bonferroni adjustment is increasingly conservative as the multiple test results become increasingly correlated. Since test results one percentile apart are likely to be more correlated than 10 percentiles apart, the Bonferroni adjustments are likely to be excessively conservative. We instead form "optimized" confidence intervals, searching over the same discrete grid looking for the test size closest to 5%.

The bottom right panel of Table 2 shows the results for simulations over 99 percentiles. The column labeled "CI" displays the size of the confidence interval, such that the rejection

[3] We search over twenty-five confidence intervals starting with 95.0% and increasing 0.2% until 99.8%.

rate of the correct model is as near as possible to 5% within the discrete grid we search over. In the top panel, the CAPM describes the population data. When the CAPM is also the candidate model, the length of the confidence interval that generates test size nearest to 5% is 98.4%. The rejection rate this yields is 5.2%. The next row preserves this same confidence interval, but now tests the (false) FF6 model, which does not describe the population data. This model is correctly rejected in 99.9% of our simulations. When the exercise is repeated for the distribution of alphas (as opposed to t-statistics), the optimized confidence interval is 99.4% and the test size yielded is 4.7%. When this length confidence interval is tested on the false FF6 model, the model correctly rejects 99.7% of the time.

The bottom of Panel B repeats this exercise for population alphas that are zero under the FF6 model. When the FF6 is the candidate model the optimized confidence interval is 99.4% and rejects the correct model 4.6% of the time. When this size confidence interval is preserved and the (false) CAPM is the candidate model, the model is rejected 96.7%. Repeating the same exercise with the distribution of alphas, yields a confidence interval of 99.6%, a rejection rate of the true model of 5.6% and a rejection rate of the false model of 90.4%.

Taken as a whole, the results in Table 2 show some gains in test power when confidence intervals are optimized to generate a test size near 5%. The test power when the more conservative Bonferroni corrections determine the confidence intervals over 9 percentiles of t-statistics is 97.5% rejecting FF6 and 83.2% rejecting the CAPM, while the optimized confidence intervals reject 99.9% and 96.7% of the time over 99 percentiles. Figure 3 shows the size-power trade-off graphically when the distribution of t-statistics are used for our tests. As the confidence interval increases, size (blue line) declines linearly. Power (red line) declines slowly at first, but more rapidly as the confidence interval widens. The dotted line shows the length of the confidence interval that generates test size of approximately 5%.

3.4 Simulation Results and Data Snooping

Our simulations create two distinct worlds. The first is a world in which the Fama and French six factor model is true. There are five additional asset pricing factors that along with the market portfolio characterize expected returns across stocks. The CAPM is not sufficient to explain expected returns and should be rejected. The second is a world in which the CAPM is true, only the market portfolio and market beta contribute to a stock's expected return. The Fama and French six factor model should be rejected.

In the multifactor setting, the CAPM is an incomplete description of asset prices. But in the CAPM setting, what is the Fama and French six factor model? Recall that we do not manipulate the Fama and French factors. The factors have CAPM alpha *in population*. Traditional asset pricing tests, such as the spanning tests, would all reject the CAPM even though it is the true model. Barillas and Shanken (2017, 2018) show that a number of asset pricing tests reduce to the spanning tests and that factor models can be tested by comparing their ex post Sharpe ratios. But in this setting the Fama and French six factor has the higher ex post Sharpe ratio, even though it is not the true model and should be rejected.

The reason the Fama and French model can be wrong and still outperform in this world is because the FF6 model is, in effect, cheating. The only way to form the FF6 factors from the individual stocks in our sample is to condition on information unknowable in advance. The FF6 factors are traded factors in the sense that they are combinations of tradeable securities, but they are untradeable in the sense that the information necessary to form them is only available ex post.

For instance, consider creating HML, which has CAPM alpha, in the CAPM world. The expected return is equal to

$$E[HML_t] = \alpha + \beta E[R_{m,t}]$$

.

The HML factor is a set a of weights, $w_{i,t}$, on each stock each period that combine returns in a way that generates the HML factor.

$$HML_t = \sum w_{i,t} R_{i,t} = \sum w_{i,t}(\beta_i R_{m,t} + \epsilon_{i,t})$$

The last equal sign follows, because the CAPM is true, so no stock has alpha. Substituting the second line into the first and using $\sum w_{i,t}\beta_i = \beta_{HML}$, gives:

$$E[HML_t] = \alpha + \beta E[R_{m,t}] = \beta_{HML} E[R_{m,t}] + E[\sum w_{i,t}\epsilon_{i,t}]$$

Creating the HML's CAPM alpha when the CAPM is true requires choosing weights so that the error terms sum to the alpha.

$$\alpha = E[\sum w_{i,t}\epsilon_{i,t}]$$

Since the error terms have ex ante expectation of zero, this requires ex post knowledge of the error terms. The HML is a traded factor, but it is not available to any investor in real time.

The FF6 factors in these simulations are impossibly good. They represent exactly the type of data snooping that is central to our motivation. It is as if researchers had searched over many possible managed portfolios and retained as factors ones with the attributes of the FF6 factors. In any set of returns, test assets can be formed on some characteristic or trading signal to generate alpha relative to a factor model. Even the true model, when tested on these assets, will be rejected. In our tests, as the simulations show, those impossibly good factors can still be rejected. Individual stocks are a natural test asset to guard against this form of data snooping.

3.5 Synthetic Alpha

The preceding simulations show that our procedure has appropriate test size, while retaining considerable power across two prominent models, the CAPM and the FF6. By using one model to generate a zero-alpha data set and another model as the candidate tested, these simulations generate a spread in alpha across stocks that is the basis to reject false models. The overall power of the test will depend on the distribution of this latent alpha. In this section, we explore the test power of our procedure across different distributions of alpha.

We start with the sample data that has been transformed to have zero alpha under the CAPM.

$$ZAP_{it} = \beta_i mktrf_t + \epsilon_{it}$$

We then generate a distribution of alpha from normal distribution with mean zero and standard deviation, σ

$$\alpha_i \sim \mathcal{N}(0, \sigma)$$

and add it to the zero alpha portfolios

$$ZAP_{it} = \alpha_i + \beta_i mktrf_t + \epsilon_{it}$$

We vary the standard deviation across a range from 0.05% to 1.00%. The larger the spread in alphas the greater power our test will have to reject false models. We repeat the simulations described above to explore the range over which we should expect reasonably powerful tests.

Figure 4 shows the results of these simulations. The power to reject false models rises quickly in the range of standard deviations of alpha from 0.25% to 0.70% from near 5% rejection rate to a 90% rejection rate. We find the power rapidly increases from over 50% to over 90% in the range from 0.60 to 0.70.

It is worth pausing to consider what a reasonable range of standard deviation of alpha across stocks is. Consider for example a world where stock returns are determined by the FF6 model:

$$R_{it} = \beta_i mktrf_t + \beta_i smb_t + \beta_i hml_t + \beta_i rmw_t + \beta_i cma_t + \beta_i mom_t + \epsilon_{it}$$

If the CAPM is the candidate model being tested, the population alpha across stocks is determined by the distribution of the stocks betas and the expected returns on the additional factors.

$$E[\alpha_i] = s_i E[smb] + h_i E[hml] + r_i E[rmw] + c_i E[cma] + m_i E[mom]$$

The standard deviation of alpha across stocks is given by

$$\sigma(\alpha) = \sqrt{E[smb]^2\sigma^2(s_i) + E[hml]^2\sigma^2(h_i) + E[rmw]^2\sigma^2(r_i) + E[cma]^2\sigma^2(c_i) + E[mom]\sigma^2(m_i)}$$

If we replace the sample realizations of average factor returns and the variance of estimated betas from our data into the equation above, the standard distribution of alpha across firms is 1.50%. Since variation in the betas is in part due to sampling error, this 1.50% overstates the true variation in betas. We can estimate the true underlying variation in betas by decomposing sample variation from the true variation (Fama and French (1997), Lewellen and Nagel (2006)).

$$\sigma^2(\hat{\beta}) = \sigma^2(\beta_{True}) + \sigma^2(e_t)$$

Under standard OLS assumptions and assuming a stable beta, the variance across estimated betas equals the true variance across betas plus the average variance of the sampling error. Using this decomposition results in a standard deviation of alphas of 0.83%, well into the range that we expect to have very high test power as shown in Figure 4.

4 Data

Our sample consists of stock returns from the Center for Research in Security Prices (CRSP). We restrict our sample to common equity (shrcd 10 or 11) from the major exchanges, New York Stock Exchange, NASDAQ and NYSE America (exchcd 1, 2 or 3). To be included in our sample a stock must have a share price and outstanding shares in the month before the return date and at least 36 non-missing returns. Our full sample starts in July of 1964 and ends in December of 2018. The sample is limited by the availability of the factor models to start in July of 1969 and end in May of 2014.[4] Since we wish to compare performance across models, we present results for all models in this subsample.

In addition to our main results, we explore two subsamples of stocks, "No Micros" and "Large Stocks." We define micro-cap stocks are defined as stocks that begin the month with less market equity that the bottom 20th percentile of the NYSE for that month, and large cap stocks as those with market equity greater than the 50th percentile of NYSE stocks (Fama and French (2008)).

We apply our simulation method to 16 risk factors proposed in the literature. Specifically, we use market (mkt), size (smb), book-to-market (hml), profitability (rmw), and investoment (cma) from Fama and French (2015), liquidity (psl) from Pástor and Stambaugh (2003), profitability (roe) and investment (ia) from Hou et al. (2015), and two composite factors (mgmt and perf) from Stambaugh and Yuan (2017). We also apply our method to six instrumented PCA factors from Kelly et al. (2019).[5]

[4]

[5] We thank Ken French, Robert Stambaugh, Lu Zhang,and Seth Pruitt for making the factors publicly available on their webpages.

5 Testing Models with Individual Stocks

In this section, we use our approach to test several of the most influential factor models in empirical asset pricing. Motivated by our simulations for each model, we show two sets of results. For the first test, we take a conservative approach. We extract alphas at each of nine deciles (from the 10th to 90th percentile) and apply the Bonferonni corrections. These tests guarantee a test size of 5% or less. The second test extracts alphas from each of ninety-nine percentiles (from the 1st to 99th percentile). We use a confidence interval of 99% for our 99 percentile tests. This number is between the optimized confidence intervals found to yield the appropriate test power in Table 2 (98.4% for CAPM and 99.4% for FF6). Based on our simulations, we expect this to generate a test size near 5% for the models tests.

Figure 5 tests the CAPM. In both the decile panel on the left and the percentile figure on the right, the average observed t-statistic is larger than the confidence interval in the middle of the distribution leading to rejection in both panels. The results suggest there are alphas (possibly latent factors) that a wide cross-section of stocks are exposed to, including the median stock in the sample.

Stocks with extreme t-statistics, the tenth and ninetieth percentile, are within the confidence interval. In traditional characteristic sorted portfolio tests most of the information is concentrated in the extreme portfolios. These stocks may have extreme loadings on a factor omitted from the pricing model, and consequently high-low extreme portfolios produce the most statistical power to reject a tested model. In these tests of individual stocks, we more commonly observe rejection towards the middle of the distribution. For individual stocks, stocks appear to have similar outcomes at the extremes as would be consistent with "luck." Extreme performances in individual stocks may have more to do with consistently surpassing (or failing to meet) investor expectations, than especially high or low cost of equity. Additionally, the confidence intervals at the tails of our distributions are often wider, since there is more uncertainty in estimating the tails of a distribution.

Next, Figure 6 shows the results for the Fama and French models. The top panels are Fama and French (1993) that adds smb (size) and hml (book-to-market) factors to the CAPM. The middle panels are Fama and French (2015) five factor models that adds rmw (profitability) and cma (investment) to the three factor model. The bottom panel, we call FF6, adds umd (momentum) to the five factor model as an updating of the popular Carhart

(1997) four factor model.

All three models are rejected across both the conservative Bonferonni tests (left column) and the optimized percentile tests (right column). The models explain more of the cross-sectional t-statistic distribution. Adding factors that explain the time-series variation of returns can have the effect of narrowing the confidence intervals by decreasing the contribution of residual variance. These narrower confidence intervals relative to the CAPM are apparent in Figure 6. Despite the narrower confidence intervals, more of the cross-sectional t-statistics are found in within the range of ordinary variation. In the left panel, both the three and five factor models are rejected in the 20th to 50th percentiles, while the six factor model is rejected at all but the top two deciles. Similar patterns are apparent in the right panel using optimized confidence intervals. This mispriced stocks are always above the intervals, suggesting stocks in the sample have alpha relative to the factors.

Next, Figure 7 shows the results for two versions of the Q-factor model. The top panel shows Hou et al. (2015) (HXZ4) which combines the market with size, profitabilty and investment factors. The bottom panel shows the model of Hou et al. (2021) (HXZ5) which adds a profitability growth factor to the four factor model. The results across the two panels are very similar. The models are rejected in both cases in a pattern similar to the FF6 factor model. Most stocks have more alpha than expected by chance, especially in the bottom of the distribution.

To streamline the results, the two models of Figure 8 are grouped by their common author, even though they are not conceptually related. The top panel shows the five factor model of Pástor and Stambaugh (2003), which adds a liquidity factor to the Carhart (1997) four factor model. The bottom panel shows the four factor mispricing model of Stambaugh and Yuan (2017). The model combines several anomalies related to management (mgmt) and performance (perf) with the market and a size factor. The top panel shows the liquidity factor does not improve much on the Fama and French models. The bottom panel shows some improvement from the mispricing model at the extremes, but the model is still rejected.

Figure 9 shows the most successful models, the IPCA factors of Kelly et al. (2019). These models use an "instrumented" form of principal components over thirty-six characteristics that have been formed into basis assets. The figure shows three versions of the model, which each have size factors. The first is the "Restricted" model of Kelly et al. (2019), which is their baseline model. The characteristics are restricted to explain average returns through common factors as opposed to allowing for characteristic mispricing. The second version is the out-of-sample version of the restricted model. The full sample model forms principal components over the entire sample, maximally explaining the variation across the basis assets. The out-of-sample version performs this in an expanding window. Lastly, the "Unrestricted" model relaxes the restricted version by allowing characteristics to explain average returns in a way unrelated to common factors.

The top two rows show that the Kelly et al. (2019) Restricted model is the only model we test that is not rejected. Both the full sample and out-of-sample versions pass our statistical tests. Neither model is rejected at any percentile. The bottom panel shows that the unrestricted version of the factors is rejected. Most stocks have higher t-statistics than would be expected by chance. The median stock has positive alpha against the unrestricted model, seemingly shifting the entire distribution of observed t-statistics up and above the confidence intervals formed under the null hypothesis.

5.1 Mispricing Measure

In this section, we suggest an intuitive approach to combine the results presented so far into a single mispricing measure. A model is accepted or rejected based on whether the observed mispricing across percentiles is within the range of the confidence interval. If the observed mispricing is far from that expected by chance, then we reject the model.

We suggest a measure of mispricing that captures how far the t-statistics of alpha are from the average of our simulations. An absolute measure of mispricing is the average of the absoluted difference at each of the nine deciles from the mean of the simulations:

$$|M| = \frac{1}{9}\sum_{i=10}^{90} |t(\alpha)_i^o - t(\alpha)_i^\mu|$$

And a squared measure of mispricing is the squared deviations from the average of the simulations:

$$M^2 = \frac{1}{9}\sum_{i=10}^{90} (t(\alpha)_i^o - t(\alpha)_i^\mu)^2$$

These measures capture the combined mispricing across the nine deciles. Additionally, by comparing to the mispricing measures calculated across the 10,000 simulations for each model, we can calculate a p-value for the extent of mispricing.[6]

Figure 10 shows this simulated p-value for the CAPM. First, we find the mean of the t-statistics at each of the nine deciles from the 10,000 simulations. Then we compare each of the simulations individually to those means at the nine deciles and find $|M|$ and M^2 for each of the simulations. Last, we calculate the mispricing measures observed in the sample and compare to the distribution of mispricing measures found in the simulations. The left panel of Figure 10 shows the absolute mispricing measure and its simulated distribution under the null hypothesis, while the right panel shows the squared mispricing measure.

Table 3 shows the resulting mispricing measures across models. The top panel shows the absolute mispricing measure and the bottom panel shows the squared mispricing measure. The models are sorted from lowest mispricing to highest mispricing and the two measures always agree on the ordering and produce similar p-values. The only model that is not rejected at the 5% level is the KPS-restricted model. As shown in Figure 9, this model has a very small absolute mispricing averaging to only one basis point across deciles. The poorest performing models, the CAPM and KPS-Unrestricted have on average 35 basis points absoluted mispricing over the nine deciles.

The KPS out-of-sample model does not perform as impressively when viewed in terms of the mispricing measure. The out-of-sample model has higher mispricing measures than all of the Fama and French style models as well as the mispricing and liquidity models. But, unlike these models, the KPS out-of-sample model is not rejected with a low p-value. The p-value of 0.08 puts this model on the margin of conventional levels of significance. This suggests that some of the success of the out-of-sample model comes from injecting additional uncertainty in the confidence intervals. A model that succeeds by blowing up the standard errors is not obviously preferable to a model that has less mispricing, but is more convincingly rejected.

The second best performing model is the Fama and French three factor model. This model has less mispricing than the Fama and French five and six factor models and the five factor model of Pástor and Stambaugh (2003) that adds momentum and liquidity to the Fama and French three factor model. Additionally, the five factor model of Hou et al. (2021) performs slightly better than its four factor counterpart. That more factors does not always

[6] We have not reported a p-value up to this point, because with multiple comparisons extreme p-values are more likely to occur by chance, distorting the usual interpretation.

generate less mispricing suggests that testing asset pricing models with individual stocks may provide a natural balance for tests that require pricing extreme characteristic portfolios or high Sharpe ratio factors.

5.2 Empirical Results Without Micro-cap Stocks

A number of the models perform poorly when tested on individual stocks. Individual stocks may be an especially high bar for factor models. To enhance robustness, factor models often curtail the weight put on stocks with very low market equity. For instance, within the small cap universe (less than 50th percentile market equity on NYSE) Fama and French (1993) value-weight stocks across characteristic sorted portfolios. So called micro-cap stocks that have a lower market equity than the 20th percentile NYSE stock are less the 3% of total market equity but account for 60% of the entire NYSE-NASDAQ-AMEX universe (Fama and French, 2008).

In this section, we drop all micro-cap stocks and reanalyze each model. Figure 11 shows a selection of models retested on this sample. Interestingly, the CAPM across the top panel is still rejected, but the next two panels show the Fama and French three and five factor models are no longer rejected. The bottom panel shows that the six factor model that includes momentum is not rejected across nine deciles, but is rejected across ninety-nine percentiles. The only instance where the two tests disagree.

Table 4 shows the results for a wider selection of models, sorted by our absolute mispricing measure. Using the mispricing measures, several models are not rejected at the 5% level of significance, including the Fama and French (2015) five factor model, the Fama and French (1993) three factor model, the Hou et al. (2015) four factor model, the Stambaugh and Yuan (2017) four factor model and the Kelly et al. (2019) out-of-sample model. Again, it is apparent that adding factors does not necessarily improve model performance. The Pástor and Stambaugh (2003) model adds momentum and liquidity factors to the Fama and French three factor model but has more mispricing and is rejected at a p-value of 0.00. Adding a momentum factor to the Fama and French five factor model also worsens performance for the Fama and French six factor model.

The overall best performing model under both absolute and squared mispricing is the Fama and French five factor model. The two measures generally agree in the rank order of the models, though small differences in the measures lead to the Hou, Xue, and Zhang four factor model falling from second to fourth under the squared mispricing measure.

The last two columns report the results for our two tests across deciles and percentiles. The second to last column shows the more conservative Bonferroni test over nine deciles and the last column shows the optimized confidence intervals over 99 percentiles. The two tests agree for all but the Fama and French six factor model. Only the KPS-Restricted model is rejected by the mispricing measures, but not by the percentile tests.

6 A Data Snooped Model

In order to further distinguish our tests from other asset pricing tests, we consider a model that we think is quite likely to be data snooped. We are confident it is data snooped, because we have done the snooping. Hou et al. (2020) replicate 452 different anomalies.[7] We take seven anomalies notable for how exceptional their Sharpe ratios are. From these characteristics, we build seven factors from value-weighted, decile sorts on each characteristic. We take the extreme high and low return portfolios and generate high minus low hedge portfolios from each characteristic.

The seven characteristics (along with their resulting monthly Sharpe ratios and the t-statistics on whether the factor returns are different than zero) are cumulative abnormal stock returns around earnings announcements (Abr1 0.21, 4.81), change in analysts earnings forecast (dEf1 0.21, 4.71), twelve-month industry lead-lag effect in prior returns (Ilr12 0.15, 3.45), twelve-month quarterly earnings to price (Epq12 0.10, 2.21), change in net operating assets (dNoa 0.11, 2.59), four-quarter change in return on equity (dRoe1 0.23, 5.23), and seasonality (R[2,5]a 0.18,4.19). We combine these seven factors with the market factor and test the resulting model on individual stocks.

Figure 12 shows that this model can be rejected using our method. Even though a mean-variance efficient combination of the resulting factor model has an extremely high Sharpe ratio (almost three times that of the market), that would help it in many asset pricing tests (Barillas and Shanken, 2017, 2018), our test can distinguish it from the true model. The average alphas are above the confidence intervals, suggesting that many stocks load

[7] We are grateful the authors have made the resulting test portfolios available at http://global-q.org/testingportfolios.html.

negatively on the factors.

It is not unique to our tests that a high Sharpe ratio factor model can still fail to price test assets.[8] But, since both generation of traditional factor models and test assets require strong assumptions about how characteristics relate to expected returns, individual stocks have the advantageous feature of avoiding this predicament entirely. There may well be instances where characteristic sorted portfolios have the advantage of focusing a test on certain hard to price relationships across stocks, and in doing so increase test power, but if the future evolutions of factor models are more closely approaching the true mean-variance frontier, improvement on these tests should not come at the expense of pricing individual stocks. Individual stocks can provide some counterbalance to explaining patterns in the data that are only apparent after that data has been thoroughly examined.

7 IPCA Models

The IPCA models of Kelly et al. (2019) have a structural difference from the other asset pricing models considered. Rather than sort stocks into value-weighted portfolios, stocks are first sorted into basis assets long each individual characteristic, then factors are extracted from these basis assets. Because their main unit of analysis is individual stocks, these basis assets are more like equal-weighted portfolios than value-weighted. They take advantage of characteristics spread across the whole cross-section and not just across large stocks. The IPCA model also differs in that it explores a much larger cross-section of "anomalies." The procedure uses 36 firm characteristics to form six factors, while the other models use sorts over only six characteristics.

In this section, we explore the net effect of these differences in approach to examine where the superior performance of the IPCA factors in individual stocks stems from. We create IPCA factors using the Kelly et al. (2019) methodology with a smaller subset of characteristics to see if the out-performance is driven by the structure of IPCA or the additional firm characteristics.[9] We reproduce the three versions of the Kelly et al. (2019) models, restricted,

[8] As Barillas and Shanken (2017) put it, "There is certainly no guarantee that the model identified as best is a good model. To address this issue, which entails evaluation of the overall performance of the model, all information about the pricing of excluded factors."

[9] We are grateful to Seth Pruitt for making the replication code available on his website `https://sethpruitt.net/research/`.

unrestricted and out-of-sample, but with only a subset of the characteristics. We use a six characteristic version that includes beta, size, book-to-market, investment, profitability and momentum. We also use a five characteristic version that omits beta and instead augments IPCA factors with the market factor.

Table 5 shows the mispricing measures for these models. Interestingly, the IPCA models perform better without the additional characteristics, especially out-of-sample. The KPS out-of-sample model with six characteristics has an absolute mispricing measure of 0.06 basis points and is not rejected with a p-value of 0.42. With all 36 characteristics the mispricing measure is 0.26 basis points and a p-value of 0.08. The IPCA model with five characteristics augmented with the market factor has the lowest mispricing measures and is not rejected with p-values of 0.16 and 0.24. The highest mispricing measures are produced by the IPCA models with thirty-six characteristics, though only the Unrestricted version is rejected. In summary, the IPCA structure is a promising way to build asset pricing models, especially if pricing the entire cross-section of stocks is the goal, but the method actually performs better when a smaller, more selective set of firm characteristics is used.

8 Conclusion

We test factor models over the cross-section of individual stocks. Using innovations emerging from the fund performance literature (Kosowski et al., 2006; Fama and French, 2010), we show that comparing observed alpha over percentiles of individual stocks to distributions simulated under the null hypothesis of no mispricing can generate appropriately sized tests with surprisingly high statistical power. We test several leading models on individual stocks and find that many can be rejected. If we omit the tiny, but plentiful, micro-cap stocks, the models perform better as a whole, but many can still be rejected.

When all stocks are used as test assets, the best factors are generated by the IPCA methodology of Kelly et al. (2019), but we find that the methodology performs better when only a small set of firm characteristics are used. Highlighting the difference between these tests and traditional asset pricing tests, our tests show models with more factors and higher ex post Sharpe ratios may perform worse on individual stocks. We test a model that has been deliberately data snooped with very high Sharpe ratio factors, and show it can be rejected when tested on individual stocks. Individual stocks do not require strong assumptions about the relationship between characteristics and expected returns, and consequently provide a valuable check on the evolution of factor models. If factor models are truly approaching the ex ante mean-variance efficient frontier, then better performance on traditional tests should not come at the expense of pricing individual stocks.

References

Ang, A., Liu, J., and Schwarz, K. (2020). Using stocks or portfolios in tests of factor models. *Journal of Financial and Quantitative Analysis*, 55(3):709–750.

Barillas, F. and Shanken, J. (2017). Which alpha? *The Review of Financial Studies*, 30(4):1316–1338.

Barillas, F. and Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance*, 73(2):715–754.

Barras, L. (2019). A large-scale approach for evaluating asset pricing models. *Journal of Financial Economics*, 134(3):549–569.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, 65(1):179–216.

Blume, M. E. (1970). Portfolio theory: a step toward its practical application. *The Journal of Business*, 43(2):152–173.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82.

Chaieb, I., Langlois, H., and Scaillet, O. (2021). Factors and risk premia in individual international stock returns. *Journal of Financial Economics*, 141(2):669–692.

Chen, Y., Cliff, M. T., and Zhao, H. (2017). Hedge funds: The good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis (JFQA)*, 52(3):1081–1109.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics.*

Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193.

Fama, E. F. and French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, 63(4):1653–1678.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance*, 65(5):1915–1947.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Ferson, W., Chen, Y., et al. (2019). How many good and bad funds are there, really?

Ferson, W. E., Sarkissian, S., and Simin, T. (1999). The alpha factor asset pricing model: A parable. *Journal of Financial Markets*, 2(1):49–68.

Ferson, W. E., Sarkissian, S., and Simin, T. T. (2003). Spurious regressions in financial economics? *The Journal of Finance*, 58(4):1393–1413.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica*, 84(3):985–1046.

Gagliardini, P., Ossola, E., and Scaillet, O. (2019). A diagnostic criterion for approximate factor structure. *Journal of Econometrics*, 212(2):503–521.

Gagliardinia, P., Ossolac, E., and Scailletd, O. (2020). Estimation of large dimensional conditional factor models in. *Handbook of Econometrics*, page 219.

Giglio, S., Liao, Y., and Xiu, D. (2019). Thousands of alpha tests. *Chicago Booth Research Paper*, (18-09):2018–16.

Harvey, C. R. and Liu, Y. (2018). Detecting repeatable performance. *The Review of Financial Studies*, 31(7):2499–2552.

Harvey, C. R. and Liu, Y. (2019). Lucky factors. *Available at SSRN 2528780.*

Harvey, C. R. and Liu, Y. (2020). Luck versus skill in the cross-section of mutual fund returns: Reexamining the evidence. *Available at SSRN 3623537.*

Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hou, K., Mo, H., Xue, C., and Zhang, L. (2021). An augmented q-factor model with expected growth. *Review of Finance*, 25(1):1–41.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

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

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., and Wang, J. (2019). Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. *Journal of Financial Economics*, 133(2):273–298.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, 61(6):2551–2595.

Lewellen, J. and Nagel, S. (2006). The conditional capm does not explain asset-pricing anomalies. *Journal of Financial Economics*, 82(2):289–314.

Lo, A. W. and MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, 3(3):431–467.

Pástor, L. and Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685.

Raponi, V., Robotti, C., and Zaffaroni, P. (2020). Testing beta-pricing models using large cross-sections. *The Review of Financial Studies*, 33(6):2796–2842.

Roll, R. and Ross, S. A. (1984). A critical reexamination of the empirical evidence on the arbitrage pricing theory: A reply. *The Journal of Finance*, 39(2):347–350.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, 30(4):1270–1315.

Table 1: Cross-section of Population Data for Simulations

We generate a population of data by subtracting estimated alpha from sample data. This table summarizes our original sample data as well as the adjusted data that has zero alpha under either the CAPM or the Fama French six factor model (FF6).

Panel A shows the data adjusted to have zero alpha under the CAPM. The second, third and fourth columns show the R-squareds, CAPM alphas, and CAPM t-statistics of time-series regressions of stock returns on the excess market return for each stock in our sample summarized at nine percentiles. These CAPM alphas are subtracted from each stock to create a pseudo population that has zero CAPM alpha. The fifth and sixth columns show the FF6 alphas and FF6 t-statistics on these adjusted returns.

Panel B shows the data adjusted to have zero alpha under the FF6. The second, third and fourth columns show the R-squareds, FF6 alphas, and FF6 t-statistics of time-series regressions of stock returns on the six Fama and French factors for each stock in our sample summarized at nine percentiles. These FF6 alphas are subtracted from each stock to create a pseudo population that has zero FF6 alpha. The fifth and sixth columns show the CAPM alphas and CAPM t-statistics on these adjusted returns.

Panel A: CAPM

				$\alpha_{CAPM} = 0$	
Percentile	R^2	α	t	α_{FF6}	t_{FF6}
1	0.01%	-6.74	-2.53	-5.89	-1.77
5	0.4%	-3.77	-1.61	-2.78	-1.17
10	1.1%	-2.38	-1.14	-1.71	-0.85
25	4.1%	-0.59	-0.38	-0.57	-0.38
50	9.9%	0.39	0.37	0.10	0.09
75	18.2%	1.03	1.01	0.59	0.61
90	27.4%	1.87	1.57	1.23	1.07
95	32.6%	2.54	1.90	1.84	1.33
99	41.9%	4.57	2.55	3.64	1.82

Panel B: FF6

				$\alpha_{FF6} = 0$	
Percentile	R^2	α	t	α_{CAPM}	t_{CAPM}
1	2.9%	-7.26	-2.52	-3.64	-1.87
5	5.7%	-3.73	-1.63	-1.84	-1.34
10	7.9%	-2.28	-1.20	-1.23	-1.06
25	13.3%	-0.65	-0.48	-0.59	-0.58
50	21.3%	0.23	0.21	-0.10	-0.08
75	31.0%	1.11	0.86	0.57	0.35
90	40.1%	2.37	1.42	1.71	0.75
95	45.7%	3.45	1.78	2.78	1.00
99	56.7%	7.18	2.46	5.89	1.45

Table 2: Simulation Results for Size and Power

This table shows size and power of our procedure based on nine percentiles with our two methods of generating confidence intervals. The top sub-panel shows size and power of our procedure when the population is the CAPM and the bottom sub-panel shows size and power of our procedure when population is generated under the FF6 model.

For the top sub-panel, we first adjust the sample data so that the CAPM alpha is zero for all stocks. This creates a new pseudo population in which we know the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 654 months and retaining the entire cross-section of stocks and factors from each month. This creates a new sample drawn from a known population. Next, to calculate observed alphas for CAPM and FF6, we run the following time-series regressions across each stock.

$R_{it} = \alpha_i + \beta_i mktrf_t + \epsilon_{it}$

$R_{it} = \alpha_i + \beta_i mktrf_t + \beta_i SMB_t + \beta_i HML_t + \beta_i RMW_t + \beta_i CMA_t + \beta_i MOM_t + \epsilon_{it}$

To calculate size and power, we first subtract the CAPM estimated alphas from the sample of stock returns creating a new pseudo population, in which, the tested model is true. We re-sample from this pseudo population to generate 1000 samples to construct our confidence intervals. We reject a model if an observed alpha is outside the confidence interval. In the last step, we calculate size (power) for each confidence interval if the CAPM (FF6) observed alpha is outside the confidence interval.

We repeat this excercise allowing both CAPM and FF6 to be null hypothesis and candidate (four combinations) and for both alphas and t-statistics.

Panel A: Bonferonni Confidence Intervals

Model	Stat	Test	Rejection
			Pop: $\alpha_{CAPM} = 0$
CAPM	t	Size	1.5%
FF6	t	Power	97.5%
CAPM	α	Size	3.2%
FF6	α	Power	98.8%
			Pop: $\alpha_{FF6} = 0$
FF6	t	Size	3.3%
CAPM	t	Power	83.2%
FF6	α	Size	4.9%
CAPM	α	Power	78.8%

Panel B: Optimized Confidence Intervals

			9 Deciles		**99 Percentiles**	
H1	Stat	Test	CI	Rejection	CI	Rejection
				Pop: $\alpha_{CAPM} = 0$		
CAPM	t	Size	97.8%	5.1%	98.4%	5.2%
FF6	t	Power	97.8%	99.9%	98.4%	99.9%
CAPM	α	Size	99%	4.7%	99.4%	4.7%
FF6	α	Power	99%	99.9%	99.4%	99.7%
				Pop: $\alpha_{FF6} = 0$		
FF6	t	Size	99%	4.9%	99.4%	4.6%
CAPM	t	Power	99%	91.0%	99.4%	96.7%
FF6	α	Size	99.4%	5.1%	99.6%	5.6%
CAPM	α	Power	99.4%	81.2%	99.6%	90.4%

Table 3: Mispricing : All Stocks

This table shows our two measures of mispricing and a p-value based on 10,000 simulations under the null hypothesis of no mispricing. Each measure is the average mispricing across nine deciles (10, 20, 30, ..., 90).

Absolute Mispricing:

$|M|= \frac{1}{9}\sum_{i=10}^{90}|t(\alpha)_i^o - t(\alpha)_i^\mu|$

Squared Mispricing:

$M^2 = \frac{1}{9}\sum_{i=10}^{90}(t(\alpha)_i^o - t(\alpha)_i^\mu)^2$

To create each measure, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10$^{\text{th}}$, 20$^{\text{th}}$,..., 90$^{\text{th}}$) or 99 percentiles (1$^{\text{st}}$,2$^{\text{nd}}$,...99$^{\text{th}}$). Then, we adjust the sample data, so that the factor model alpha is zero for all stocks. This creates a new pseudo population in which the factor model describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months and retaining the entire cross-section of stocks and factors from each month and estimate the alphas by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times.

We calculate the mispricing measures from the mean t-statistic of our 10,000 simulations. We use the distribution of mispricing over the 10,000 simulations to report an empirical p-value.

Model	$\|M\|$	**P-Value**	M^2	**P-Value**
KPS-R	0.01	0.89	0.000	0.90
FF3	0.14	0.01	0.022	0.01
FF5	0.17	0.00	0.029	0.00
Stambaugh & Yuan	0.20	0.00	0.043	0.00
FF4 + Liquidity	0.23	0.00	0.056	0.00
FF6	0.23	0.00	0.057	0.00
KPS-OOS	0.26	0.08	0.068	0.09
HXZ4	0.28	0.00	0.083	0.00
HXZ5	0.30	0.00	0.094	0.00
CAPM	0.35	0.00	0.122	0.00
KPS-U	0.35	0.00	0.134	0.00

Table 4: Mispricing : No Micro Stocks

This table shows results when our sample is reduced by eliminating micro-cap stocks (market equity below 20th percentile of NYSE). In the first four columns, the table shows our two measures of mispricing and a p-value for this sample. The last two columns show the main hypothesis tests at each of nine deciles (with the Bonferonni correction) or ninety-nine percentiles (with our optimized confidence intervals).

For each test, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the factor model alpha is zero for all stocks. This creates a new pseudo population in which the factor model describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months and retaining the entire cross-section of stocks and factors from each month and estimate the alphas by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times.

We calculate the mispricing measures from the mean t-statistic of our 10,000 simulations. We use the distribution of mispricing over the 10,000 simulations to report an empirical p-value.

Model	$\lvert M \rvert$	**P-Value**	M^2	**P-Value**	**Reject(9)**	**Reject(99)**
FF5	0.02	0.87	0.000	0.89	No	No
HXZ4	0.09	0.16	0.010	0.18	No	No
FF3	0.09	0.06	0.009	0.08	No	No
FF6	0.09	0.02	0.009	0.04	No	Yes
Stambaugh & Yuan	0.10	0.07	0.012	0.09	No	No
KPS-R	0.13	0.02	0.020	0.03	No	No
HXZ5	0.16	0.00	0.027	0.01	Yes	Yes
FF4 + Liquidity	0.19	0.00	0.038	0.00	Yes	Yes
KPS-OOS	0.25	0.20	0.070	0.19	No	No
CAPM	0.30	0.00	0.092	0.00	Yes	Yes
KPS-U	0.62	0.00	0.403	0.00	Yes	Yes

Table 5: Mispricing : KPS Models

This table shows our two measures of mispricing for different IPCA models (Kelly et al., 2019). Models with all characteristics include the original 36 firm characteristics. Five characteristics include size, book-to-market, profitability, investment and momentum formed into five IPCA factors and augmented with the market factor. Six characteristics adds beta as a characteristic and includes only IPCA factors. "OOS" designates that the model is formed out-of-sample in an expanding window fashion.

Absolute Mispricing:

$|M| = \frac{1}{9}\sum_{i=10}^{90} |t(\alpha)_i^o - t(\alpha)_i^\mu|$

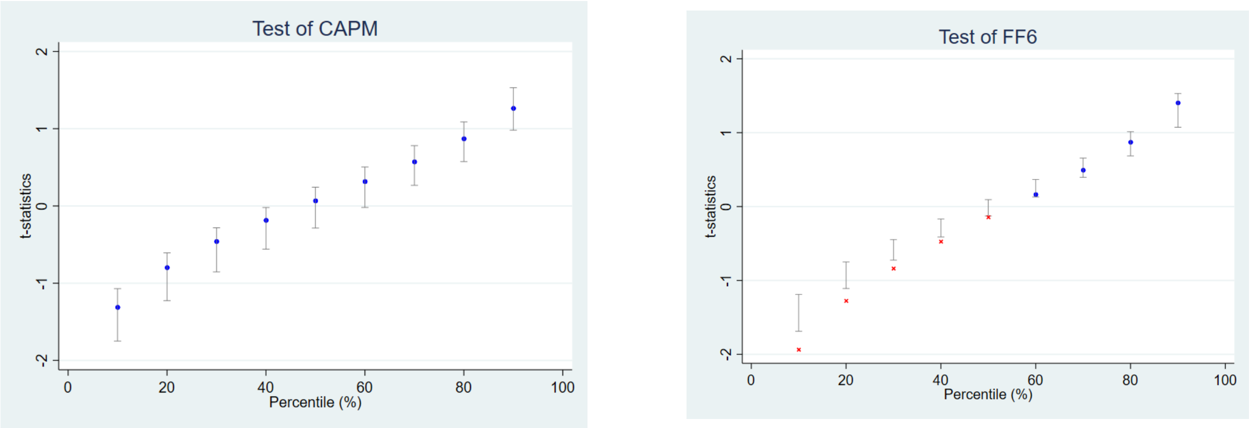
Squared Mispricing:

$M^2 = \frac{1}{9}\sum_{i=10}^{90} (t(\alpha)_i^o - t(\alpha)_i^\mu)^2$

To create each measure, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the factor model alpha is zero for all stocks. This creates a new pseudo population in which the factor model describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months and retaining the entire cross-section of stocks and factors from each month and estimate the alphas by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times.

We calculate the mispricing measures from the mean t-statistic of our 10,000 simulations. We use the distribution of mispricing over the 10,000 simulations to report an empirical p-value.

	Test Statistic		**P-Value**	
KPS Models	$\|M\|$	M^2	$\|M\|$	M^2
KPS-Restricted (All characteristics)	0.01	0.000	0.89	0.90
Market + KPS (Five characteristics)	0.05	0.003	0.16	0.24
KPS-OOS (Six characteristics)	0.06	0.004	0.42	0.49
KPS-Restricted (Six characteristics)	0.07	0.005	0.05	0.10
KPS-Unrestricted (Six characteristics)	0.11	0.013	0.00	0.00
KPS-OOS (All characteristics)	0.26	0.068	0.08	0.09
KPS-Unrestricted (All characteristics)	0.35	0.134	0.00	0.00



(a) CAPM

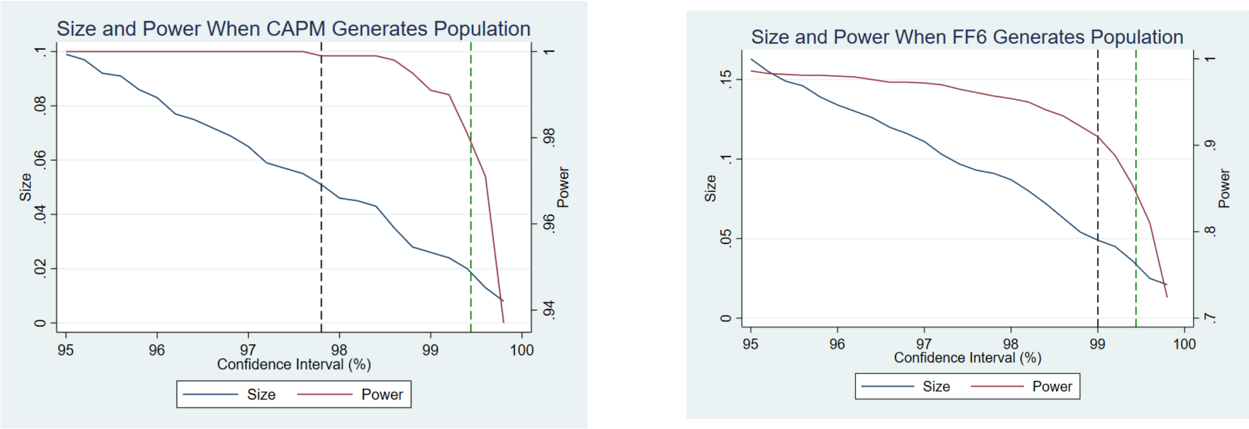
(b) FF6

Figure 1: One Simulation Run When the CAPM Describes the Population

This figure shows the first simulation run when the CAPM describes the population. The left panel shows our test of the CAPM and the right panel shows our test of the FF6 model. When the observed alphas are within the confidence intervals, we fail to reject the model (represented with a (blue) dot). When the sample observed alphas are outside the confidence intervals, we reject the model (represented with a (red) X).

First, we adjust the sample data so that the CAPM alpha is zero for all stocks. This creates a new pseudo population in which we know the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 654 months and retaining the entire cross-section of stocks and factors from each month. This creates a new sample drawn from a known population. We then test two models, CAPM and FF6. First, we estimate each model using N time-series regressions across each stock. From the estimated alpha, we retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}). This observed alpha distribution is represented with (blue) dots and (red) x's. We then subtract these alphas from the sample of stock returns creating a new pseudo population, in which, the tested model is true. We re-sample from this pseudo population to generate 1000 samples and retain the 0.28 and 99.72 percentiles to generate the Bonferroni adjusted confidence intervals. If an observed alpha is outside the confidence interval we reject the model.

Since the CAPM describes the population, a rejection represents an incorrect rejection of the null hypothesis. Since the FF6 does not describe the population, rejection represents a success, correctly rejecting a false null hypothesis. The left panel shows the results of one simulation run of the CAPM tested on a sample drawn from a CAPM population. Then, after repeating this simulation several times, the rejection rate of the CAPM gives the overall test size, and the rejection of the FF6 gives the overall test power.



(a) CAPM

(b) FF6

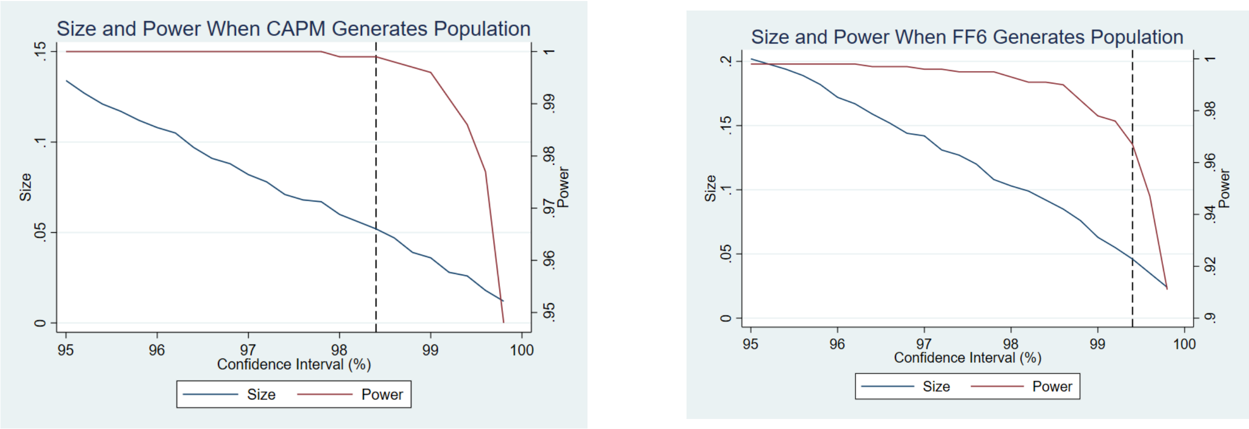
Figure 2: Size-Power Trade-off

This figure shows size (blue line) and power (red line) of our procedure based on nine percentiles. The left panel shows size and power of our procedure when population is generated under the CAPM and the right panel shows size and power of our procedure when population is generated under the FF6 model. The green dotted line is the Bonferroni adjusted confidence interval. The black dotted line is the optimized confidence interval. For the left panel, we first adjust the sample data so that the CAPM alpha is zero for all stocks. This creates a new pseudo population in which we know the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 654 months and retaining the entire cross-section of stocks and factors from each month. This creates a new sample drawn from a known population. Next, to calculate observed alphas for CAPM and FF6, we run the following time-series regressions across each stock.

$$R_{it} = \alpha_i + \beta_i mktrf_t + \epsilon_{it}$$

$$R_{it} = \alpha_i + \beta_i mktrf_t + \beta_i SMB_t + \beta_i HML_t + \beta_i RMW_t + \beta_i CMA_t + \beta_i MOM_t + \epsilon_{it}$$

To calculate size and power, we first subtract the CAPM estimated alphas from the sample of stock returns creating a new pseudo population, in which, the tested model is true. We re-sample from this pseudo population to generate 1000 samples to construct our confidence intervals. We reject a model if an observed alpha is outside the confidence interval. In the last step, we calculate size (power) for each confidence interval if the CAPM (FF6) observed alpha is outside the confidence interval.



(a) CAPM

(b) FF6

Figure 3: Size-Power Trade-off

This figure shows size (blue line) and power (red line) of our procedure based on ninety nine percentiles. The left panel shows size and power of our procedure when population is generated under the CAPM and the right panel shows size and power of our procedure when population is generated under the FF6 model. The black dotted line is the optimized confidence interval.

For the left panel, we first adjust the sample data so that the CAPM alpha is zero for all stocks. This creates a new pseudo population in which we know the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 654 months and retaining the entire cross-section of stocks and factors from each month. This creates a new sample drawn from a known population. Next, to calculate observed alphas for CAPM and FF6, we run the following time-series regressions across each stock.

$R_{it} = \alpha_i + \beta_i mktrf_t + \epsilon_{it}$

$R_{it} = \alpha_i + \beta_i mktrf_t + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + m_i MOM_t + \epsilon_{it}$

To calculate size and power, we first subtract the CAPM estimated alphas from the sample of stock returns creating a new pseudo population, in which, the tested model is true. We re-sample from this pseudo population to generate 1000 samples to construct our confidence intervals. We reject a model if an observed alpha is outside the confidence interval. In the last step, we calculate size (power) for each confidence interval if the CAPM (FF6) observed alpha is outside the confidence interval

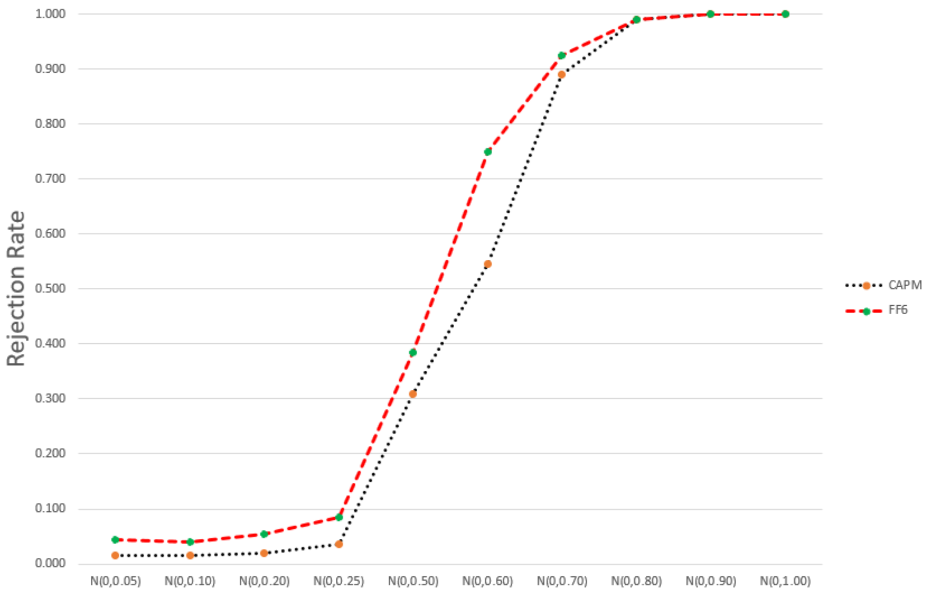
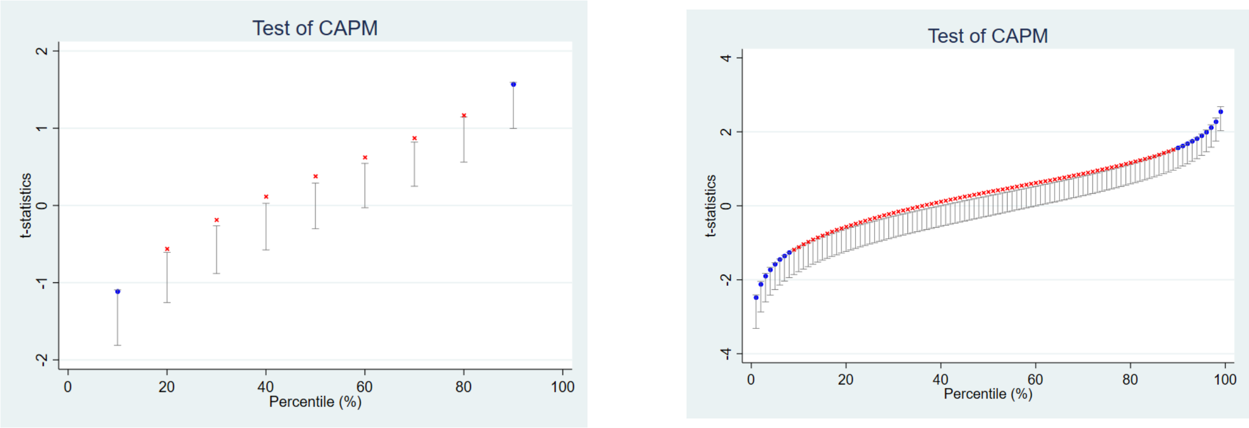


Figure 4: Rejection Frequencies of Zero-Alpha CAPM based on Different Distribution of Alpha

This figure shows the sensitivity of our procedure to different distribution of alpha. We generate synthetic alpha using normal distribution with mean zero and standard deviation, σ. We vary the standard deviation from 0.05% to 1.00%.

We first create pseudo population in which the CAPM describes expected returns on all stocks. Next, we generate normally distributed synthetic alphas and add them to our pseudo population. By doing so, we are able to test power of our procedure based on different distribution of synthetic alphas.



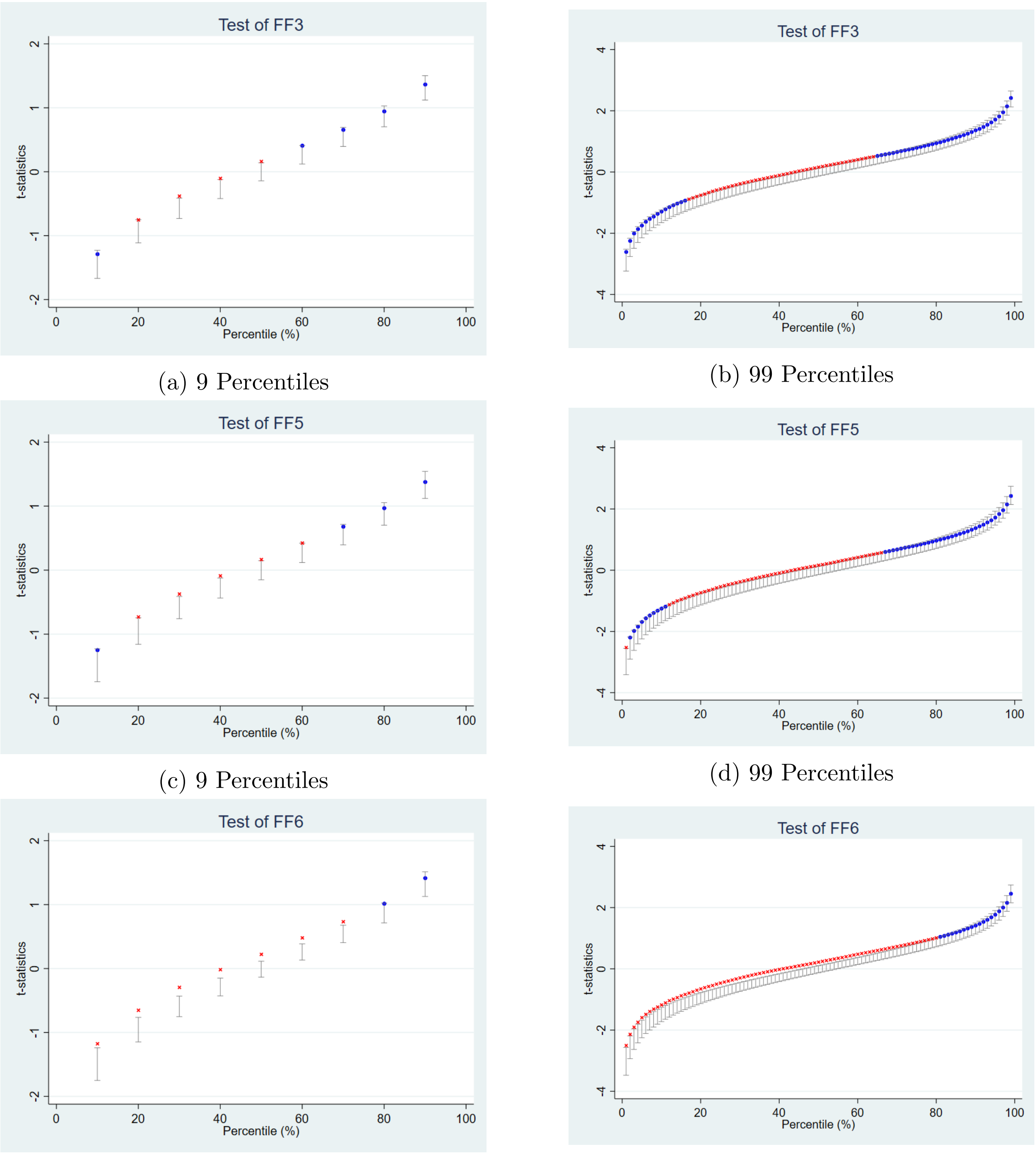
(a) 9 Percentiles

(b) 99 Percentiles

Figure 5: Testing the CAPM Using All Stocks

This figure shows the test of the CAPM for all stocks. The left panel shows the test of the CAPM for nine percentiles with Bonferroni adjusted confidence intervals. The right panel shows the test of CAPM for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



(a) 9 Percentiles

(b) 99 Percentiles

(c) 9 Percentiles

(d) 99 Percentiles

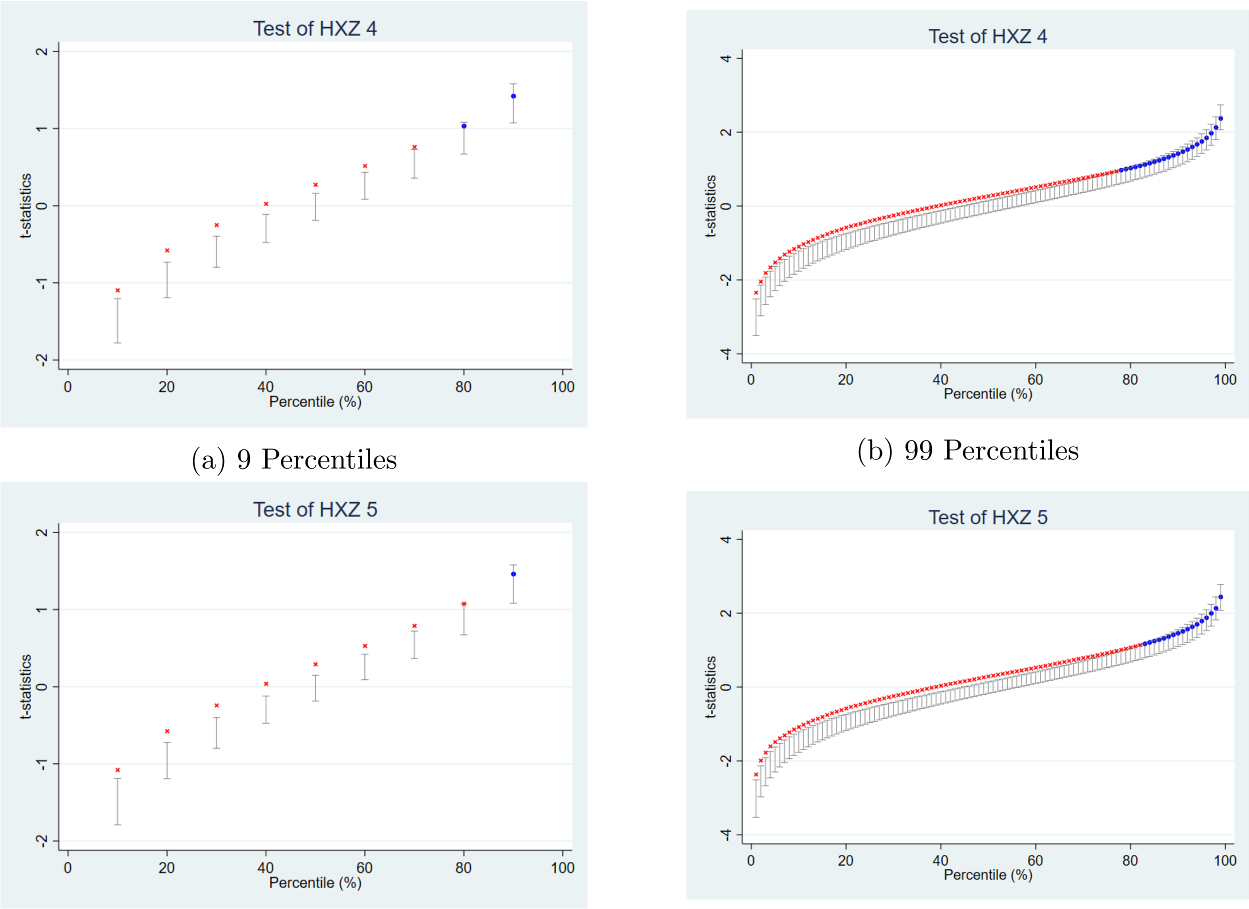
(e) 9 Percentiles

(f) 99 Percentiles

Figure 6: Testing the Fama and French Models Using All Stocks

This figure shows the test of three Fama and French models for all stocks. All models include the market. The FF3 model adds smb (size) and hml (value), the FF5 model adds rmw (profitability and cma (investment), and the FF6 adds umd (momentum). The left column shows the test of the models for nine percentiles with Bonferroni adjusted confidence intervals. The right column shows the test of the models for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



(a) 9 Percentiles

(b) 99 Percentiles

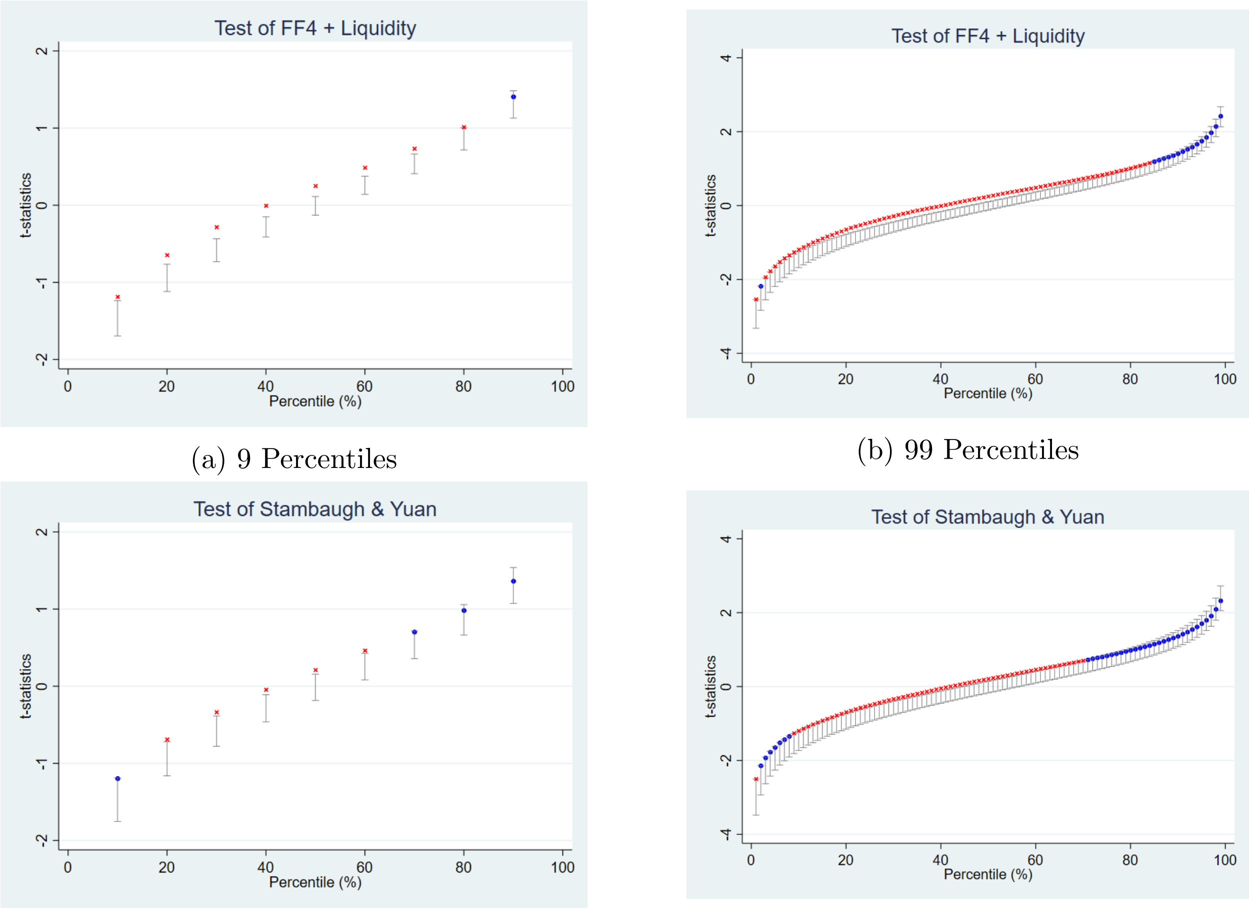
(c) 9 Percentiles

(d) 99 Percentiles

Figure 7: Testing the HXZ Using All Stocks

This figure shows the test of the two Hou et al. (2015, 2021) models for all stocks. HXZ4 includes the market, size, investment and profitability factors. HXZ5 adds profitability growth. The left column shows the test of the models for nine percentiles with Bonferroni adjusted confidence intervals. The right column shows the test of the models for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10th, 20th,..., 90th) or 99 percentiles (1st,2nd,...99th). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



(a) 9 Percentiles

(b) 99 Percentiles

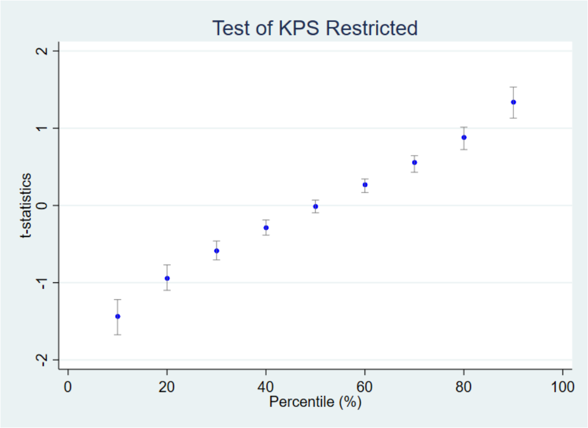
(c) 9 Percentiles

(d) 99 Percentiles

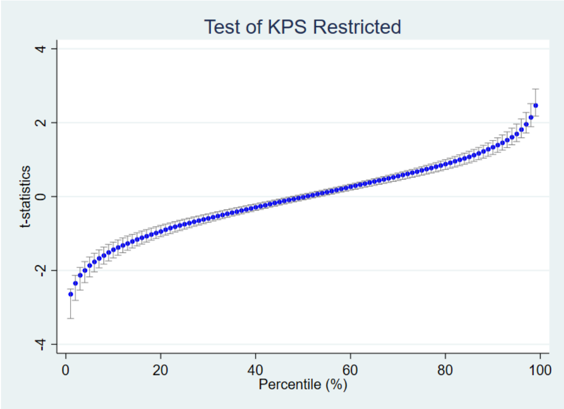
Figure 8: Testing the Stambaugh Models Using All Stocks

This figure shows the test of the FF4 plus liquidity model for all stocks and the Mispricing model. The left column shows the test of the models for nine percentiles with Bonferroni adjusted confidence intervals. The right column shows the test of the models for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

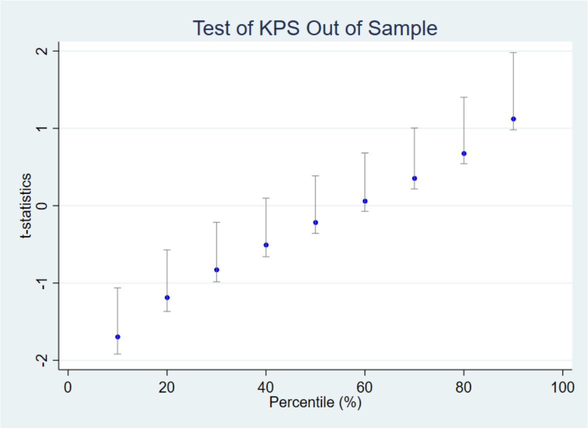
To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10$^{\text{th}}$, 20$^{\text{th}}$,..., 90$^{\text{th}}$) or 99 percentiles (1$^{\text{st}}$,2$^{\text{nd}}$,...99$^{\text{th}}$). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



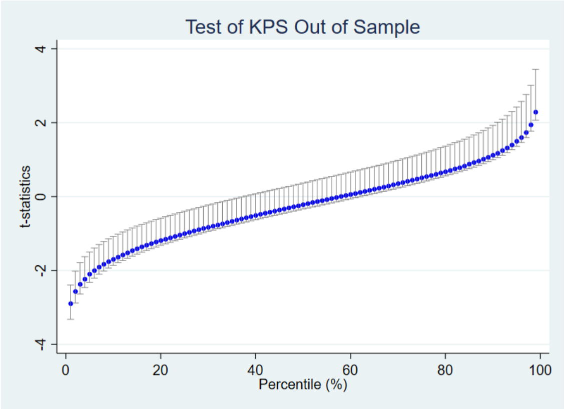
(a) 9 Percentiles



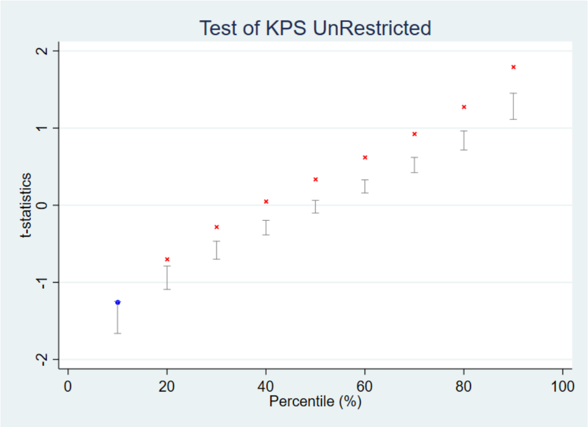
(b) 99 Percentiles



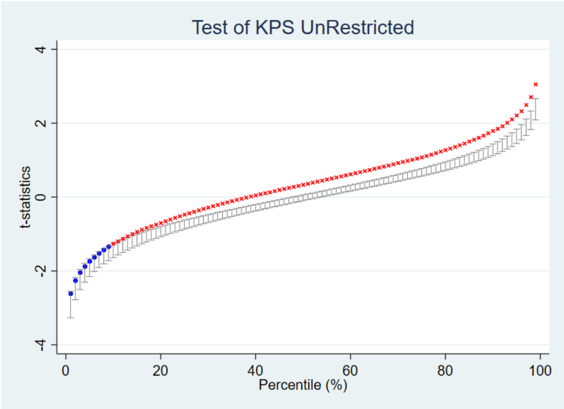
(c) 9 Percentiles



(d) 99 Percentiles



(e) 9 Percentiles

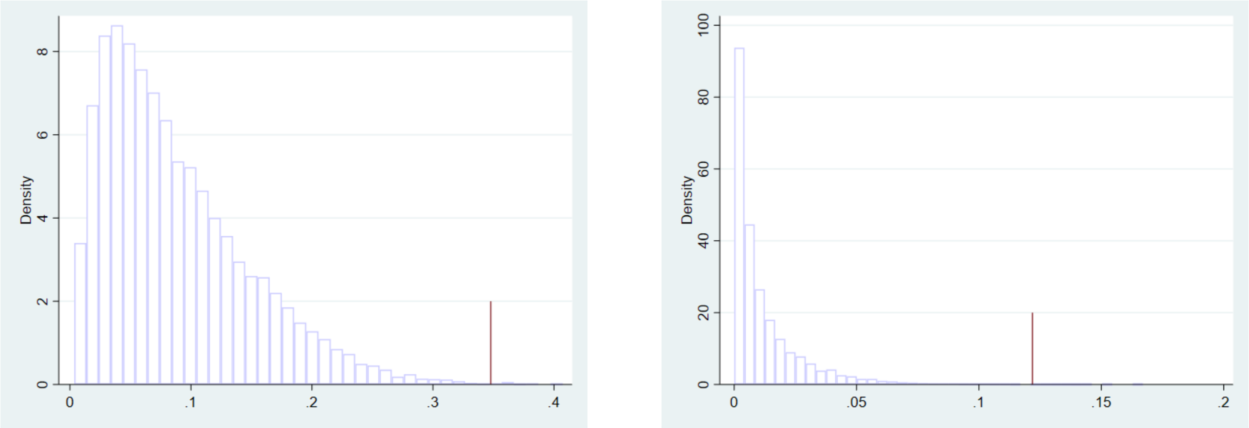


(f) 99 Percentiles

Figure 9: Testing the KPS Using All Stocks

This figure shows the test of the three Kelly et al. (2019) models for all stocks. The top panel is the KPS Restricted model that restricts characteristics to explain returns through common factors. The middle panel is the same KPS restricted model with estimates formed exclusively out of sample. The bottom panel is the Unrestricted version that does not restrict characteristics to contribute to returns through common factors. The left column shows the test of the models for nine percentiles with Bonferroni adjusted confidence intervals. The right column shows the test of the models for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10$^{\text{th}}$, 20$^{\text{th}}$,..., 90$^{\text{th}}$) or 99 percentiles (1$^{\text{st}}$,2$^{\text{nd}}$,...99$^{\text{th}}$). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



(a) Absolute difference

(b) Squared difference

Figure 10: All Stocks

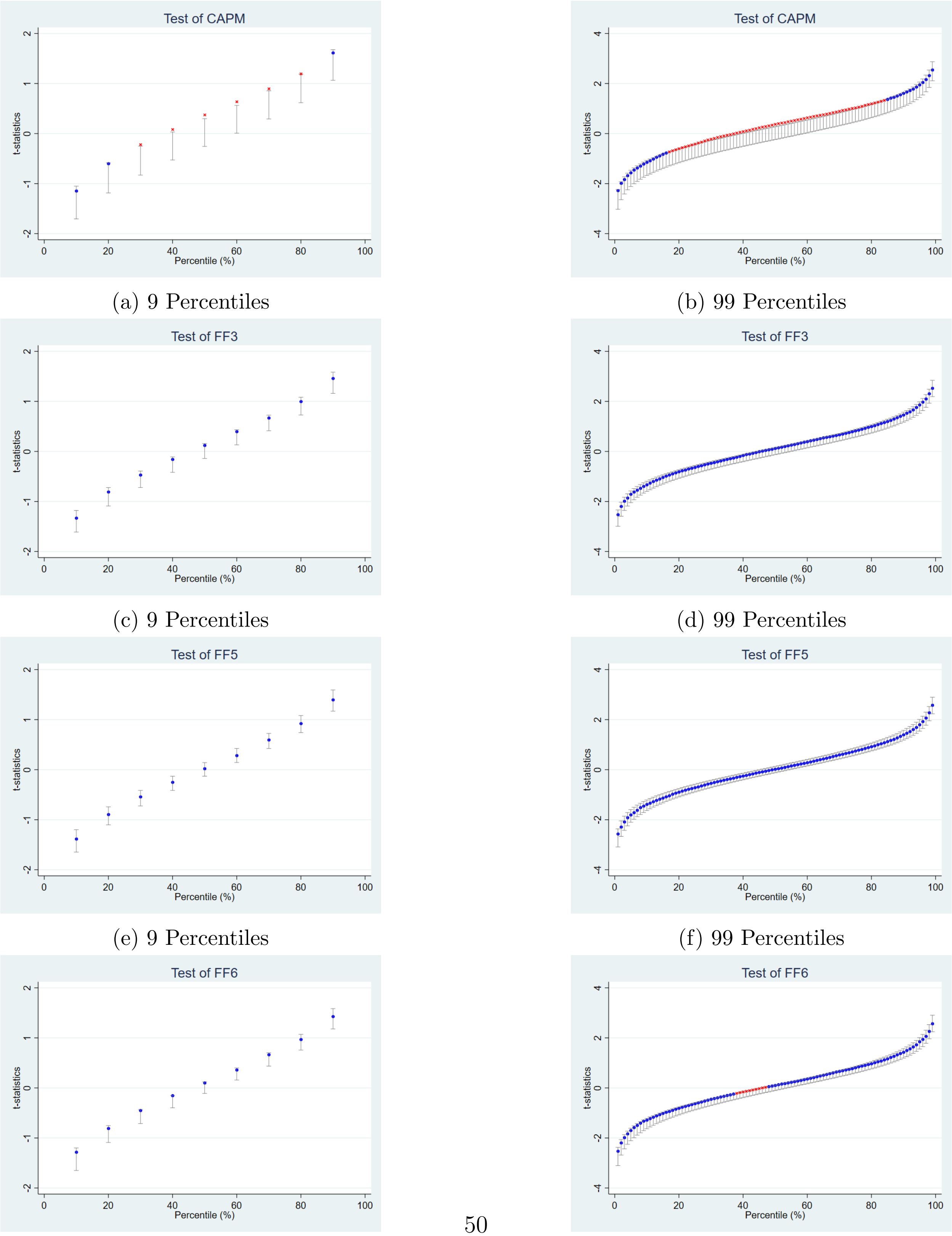
This figure shows our two measures of mispricing (red line) and histogram bins of realized mispricing from 10,000 simulations under the null hypothesis of no mispricing (blue bins). We construct p-values from the distribution of mispricing across the simulations. Each measure is the average mispricing across nine deciles (10, 20, 30, ..., 90).

Absolute Mispricing:

$|M|= \frac{1}{9}\sum_{i=10}^{90}|t(\alpha)_i^o - t(\alpha)_i^\mu|$

Squared Mispricing:

$M^2 = \frac{1}{9}\sum_{i=10}^{90}(t(\alpha)_i^o - t(\alpha)_i^\mu)^2$



(a) 9 Percentiles

(b) 99 Percentiles

(c) 9 Percentiles

(d) 99 Percentiles

(e) 9 Percentiles

(f) 99 Percentiles

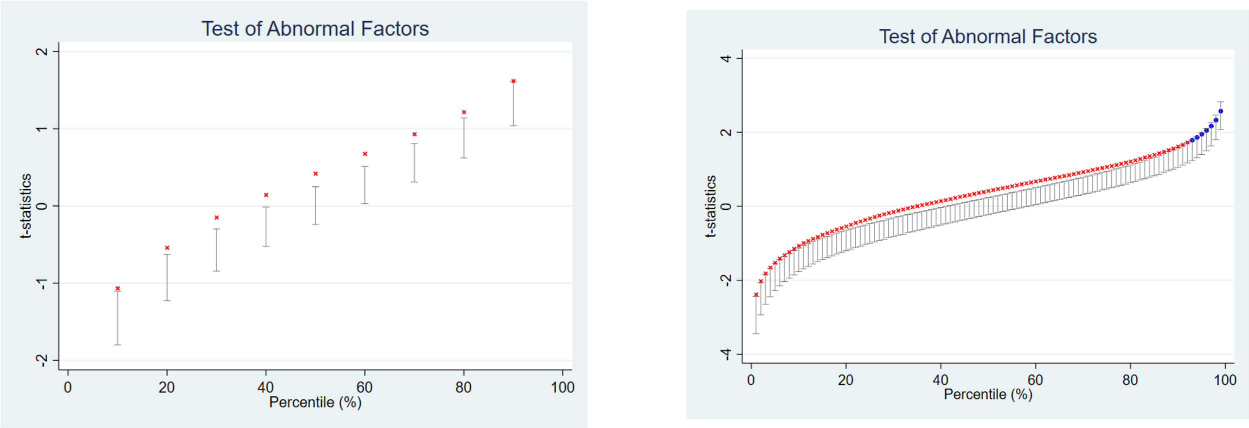
(g) 9 Percentiles

(h) 99 Percentiles

Figure 11: Testing the CAPM and Fama and French Models Dropping Micro-cap Stocks

This figure shows the test of three Fama and French models model for large and small cap stocks, dropping the micro-cap stocks. All models include the market. The FF3 model adds smb (size) and hml (value), the FF5 model adds rmw (profitability and cma (investment), and the FF6 adds umd (momentum). The left column shows the test of the models for nine percentiles with Bonferroni adjusted confidence intervals. The right column shows the test of the models for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.



(a) 9 Percentiles

(b) 99 Percentiles

Figure 12: Testing Abnormal Factors Using All Stocks

This figure shows the test of the Abnormal Factors model for all stocks. The left panel shows the test of the Abnormal Factors model for nine percentiles with Bonferroni adjusted confidence intervals. The right panel shows the test of Abnormal Factors model for ninety-nine percentiles with our optimized confidence intervals. The sample observed alphas are represented with (blue) dots and (red) x's for with red x's denoting rejections for lying outside the confidence intervals.

To perform our tests, we first use the sample data to estimate the sample observed alpha across all stocks. We retain the nine percentiles (10^{th}, 20^{th},..., 90^{th}) or 99 percentiles (1^{st},2^{nd},...99^{th}). Then, we adjust the sample data, so that the CAPM alpha is zero for all stocks. This creates a new pseudo population, in which the CAPM describes expected returns on all stocks. We then bootstrap by randomly drawing 539 months with replacement from this population and retain the entire cross-section of stocks and factors from each month. We estimate the alphas and corresponding t-statistics by running time-series regressions for each stock. We retain the percentiles for each new sample generated from this pseudo population. We repeat this procedure 10,000 times to calculate the confidence intervals at each percentile.