NBER WORKING PAPER SERIES

A PANEL REGRESSION APPROACH TO HOLDINGS-BASED FUND PERFORMANCE MEASURES

Wayne E. Ferson
Junbo L. Wang

Working Paper 28238
http://www.nber.org/papers/w28238

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2020

This paper is developed from and supersedes Wang's PhD thesis at the University of Southern California, entitled "Can weight-based measures distinguish between informed and uninformed fund managers." We are grateful to Hank Bessembinder, Wayne Chang, Chris Jones, Min S. Kim, David Solomon, Scott Joslin, Yan Liu, Jinchi Lv, Lucian Taylor Russ Wermers, the editor Thierry Foucault and anonymous referees for comments. This paper has also benefitted from workshops at Boston College, Florida International University, the Universities of Georgetown, Hawaii, Laval, Massachusetts at Amherst, Southern California, South Carolina, Washington, York, the Rutgers Business School and from presentations at the 2016 Northern Finance Association meetings, the 2017 Research Affiliates Advisory Board, the 2017 Society for Financial Econometrics meetings, the 2017 NBER-NSF time-series conference, the 2017 Southern Finance meetings the 2018 American Finance Association meetings and the 2018 Citrus Finance Conference. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

A Panel Regression Approach to Holdings-based Fund Performance Measures
Wayne E. Ferson and Junbo L. Wang
NBER Working Paper No. 28238
December 2020
JEL No. G11

ABSTRACT

Portfolio performance measures using holdings data are panel regressions. The returns of a fund's stocks are regressed on its lagged portfolio weights. Stock fixed effects isolate average performance from time-series predictive ability. Control variables condition fund performance on the characteristics of the stocks held. The long term performance of average holdings drives some of the classical measures, while predictive ability drives others. A "buy-and-hold drift," where portfolio weights increase over time in the higher alpha stocks, affects performance measures. Investor flows respond to average performance net of the buy-and-hold drift.

Wayne E. Ferson
Department of Finance
and Business Economics
University of Southern California
3670 Trousdale Parkway Suite 308
Los Angeles, CA 90089-0804
and NBER
ferson@marshall.usc.edu

Junbo L. Wang
School of Business
Louisiana State University
Baton Rouge, Louisiana
junbowang@lsu.edu

A data appendix is available at http://www.nber.org/data-appendix/w28238

Introduction

Estimates of holdings-based fund performance measures are viewed as predictive panel regressions. For each fund, the future excess returns of the stocks held are regressed on the lagged portfolio weights on the stocks. The panel regression coefficient for a fund measures its performance. An informed manager's portfolio weights should predict the stock returns with a positive coefficient (Grinblatt and Titman, 1989a). Classical performance estimates implicitly use such panel regressions, and we show that being explicit about this provides several new insights. The returns and/or weights are benchmarked or demeaned in the classical measures and the special cases depend on how this is done. We study the portfolio change measure of Grinblatt and Titman (GT, 1993), the Characteristic Selectivity measure of Daniel, Grinblatt, Titman and Wermers (DGTW, 1997), the Conditional Weight Measure of Ferson and Khang (CWM, 2002) and the stochastic discount factor measure of Ferson and Mo (FM, 2016).

Understanding these "classical" measures as special cases of a panel regression allows us to apply various approaches to clustering, fixed effects and control variables. This is important because funds may deviate substantially from common assumptions in panel regression models, such as iid model innovations. Control variables produce characteristics-based versions of conditional performance evaluation (Ferson and Schadt, 1996), but instead of conditioning on macroeconomic factors, the panel naturally asks if the performance of a fund is driven by the charactersitcs of the stocks it holds. This is a central question in studies like Carhart (1997), Wermers (2003), Zheng (1999), Sapp and Tiwari (2004), Sheng, Simutin and Zhang (2019) and Matalin-Saez (2020). We illustrate using the stocks' dividend yields as the control, and find that it removes a finding that high-yield funds perform better.

We examine the use of fixed effects in the panel regressions. Time fixed effects produce measures of performance in excess of an equally-weighted portfolio, and a common intercept has the same effect. More interestingly, stock fixed effects decompose the performance measures into the dynamic, time-series information in funds' portfolio weights and the average cross-sectional information. Time-series predictive ability (TSA) is defined as the average across stocks, of the covariances between the current portfolio weights and future stock returns. TSA reflects both factor timing and short-term security selection information. The Average Abnormal Return (AAR) is the fund abnormal return based on its average weights. The average weights are a proxy for the long term asset allocation policy (e.g., Brinson, Hood and Bebower (BHB, 1986) and Brinson, Singer and Beebower, 1991). Thus, the AAR is interpreted as a long-run measure of performance driven by fund policy or style. One of our insights is that the classical performance measures reflect both components of performance.

A fund with larger AAR on average overweights the high and underweights the low alpha stocks. Many "quant" portfolios have this feature, focusing the investment policy on cross-sectional patterns in average abnormal returns such as size, book-to-market, momentum and "fundamental" security characteristics associated with alphas in the popular models. The AAR has received little attention in the literature. However, we find that the AAR is the largest component of the abnormal performance in some of the holdings-based measures (DGTW, FM). These measures pay more attention to funds' long-term policy or style than to efforts to predict short term returns. For other measures (GT, CWM) the TSA is the dominant component, capturing mainly ability to predict returns in the short term. A panel regression view of the measures naturally isolates the TSA and AAR components.

The panel approach reveals a lagged stochastic regressor bias, similar to that described by Stambaugh (1999) and Hjalmarsson (2008), when there are stock fixed effects. We evaluate alternative ways to address the bias using simulations, including estimators from Hjalmarsson (2008, 2010) and a differenced instrumental variables approach similar to Anderson and Hsiao (1981) and Wang (2015). We find that the differenced instrumental variables and Hjalmarsson (2010) estimators work well. The previous literature provides scant information about the statistical properties of holdings-based performance measures.[2]

To illustrate AAR and TSA consider the expected stochastic discount factor alpha of a fund with portfolio weight vector $\boldsymbol{w}_t$ and weight-based excess return $r_{pt+1} = \boldsymbol{w}_t'\boldsymbol{r}_{t+1}$, where $\boldsymbol{r}_{t+1}$ is the vector of securities' returns in excess of a short-term Treasury. The scalar variable m_{t+1} is the stochastic discount factor and the unconditional expected alpha is $\alpha_p = E\{m_{t+1}\ r_{pt+1}\}$. (See Farnsworth et. al. (2002) or Ferson and Lin (2014) for applications to fund performance.) This decomposes as:

$$\alpha_p = \mathrm{Cov}\{\boldsymbol{w}_t';\ m_{t+1}\ \boldsymbol{r}_{t+1}\} + E\{m_{t+1}\ \boldsymbol{r}_{t+1}\}'\ E\{\boldsymbol{w}_t\}. \qquad (1)$$

The notation $\mathrm{Cov}\{\boldsymbol{x'};\boldsymbol{y}\}$ denotes the sum of the covariances across the elements of the vectors. The first term on the right hand side of Equation (1) is the time-series ability, or TSA. TSA reflects both short-term factor timing and selectivity ability, as explained below.[3] The second term of Equation (1) is the AAR, equal to $\boldsymbol{\alpha}'E(\mathbf{w})$, where $\boldsymbol{\alpha} = E\{m_{t+1}\ \boldsymbol{r}_{t+1}\}$ is the vector of the securities' alphas in the model. Both the TSA and the AAR depend on the model that defines the securities' alphas.

If the vector of alphas was zero; that is, the benchmark was mean-variance efficient, the AAR would be zero. Some of the literature makes the assumption that the benchmark is efficient (e.g.,

[2] Kothari and Warner (2001) use simulations to examine the power of holdings-based event study measures. Jiang, Yao and Tong (2007) use simulations to study weight-based timing measures. Ferson and Khang (2002) examine conditional weight-based performance measures using the Generalized Method of Moments.

[3] Studies that measure timing and selectivity with holdings-based measures include DGTW (1997), Becker et al. (1999), Wermers (2000), Jiang et al. (2007), Kacperczyk, Nieuwerburgh and Veldkamp. (2014), FM and others.

Grinblatt and Titman, 1989a), in which case holdings-based measures capture only TSA. However, the benchmarks used in practice are not minimum variance efficient (Cremers, Petajisto and Zitzewitz, 2012) so there will generally be nonzero AAR. Modern portfolio management (e.g., Grinold and Kan, 1995) is based on the premise there are alphas for managers to exploit. Using fixed effects in the panel isolates the TSA and the AAR so they may be evaluated separately.

If a fund rebalances the portfolio to constant weights, its performance is pure AAR. Some writers advocate active rebalancing strategies, but passive strategies can also produce AAR. A phenomenon we call *buy-and-hold drift* generates AAR. Over any multiperiod evaluation, buy-and-hold weights drift towards the stocks with higher returns during the evaluation period. Because of the high correlation between average returns and alphas, the weights drift toward higher alpha stocks. (In our sample, the correlations between stocks' average returns and their unconditional alphas range from 71% to 84% across the models.) We construct a simple measure of this buy-and-hold drift, starting from a fund's initial weights, to isolate the passive component of AAR. We find that a fund's AAR is in some but not all cases driven by the passive buy-and-hold drift, and a fund's AAR is more strongly related to its tendency towards buy-and-hold trading than to its tendency towards momentum trading.

We measure the response of new money flows to the components of performance. We find that both TSA and AAR predict flows, so fund investors seem to respond to both long-term and short-term performance. However, there is no significant flow response to the passive buy-and-hold drift component of AAR. This seems consistent with the argument of Berk and Green (2004) and Berk and vanBinsbergen (2015), where funds with capital in excess of the optimal amount are expected to passively index the excess capital. We examine the persistence of performance using the TSA and AAR components and we find that the long term AAR is the better predictor of future performance.

If the AAR component of performance reflects the long term policy of the fund, it is not likely to be related to active management. We examine of cross-sections of funds sorted by standard proxies for active management and find little relation to the AARs. We illustrate the use of control variables in the panel approach, where we use the stocks' dividend yields as the control variable.

The rest of the paper is organized as follows. Section 2 introduces the holdings-based measures of performance. Section 3 shows how classical holdings-based measures of performance are panel regressions. We discuss fixed effects in the regressions and reduced-bias methods. Section 4 describes the data. Section 5 presents simulation results and Section 6 presents empirical results for active US equity mutual funds. Section 7 concludes the paper. An Appendix presents ancillary results. An Internet Appendix is also available.

2. Holdings-Based Performance Measures

2.1 A review of the measures

Denote the portfolio weights of a fund with N stocks at time *t* as $\mathbf{w}_t = [w^1{}_t, \ldots w^N{}_t]'$ and the time-*t* stock returns in excess of a Treasury bill as $\mathbf{r}_t = [r^1{}_t, \ldots r^N{}_t]'$. Holdings-based performance measures are usually described as versions of $\mathrm{cov}(\mathbf{w}_t'; \mathbf{r}_{t+1}) = \sum_i \mathrm{cov}(w^i{}_t, r^i{}_{t+1})$, the sum across stocks of the covariances between the current weights and the future stock returns. Work on holdings-based performance measures goes back to Cornell (1979) and Copeland and Mayers (1982). Grinblatt and Titman (1989a, Proposition 4) show that an agent with an informative signal about future returns and nonincreasing Rubinstein absolute risk aversion will display a positive covariance, summed across the holdings.

From the definition of covariance we can write:

$$\mathrm{cov}(\mathbf{w}_t';\mathbf{r}_{t+1}) = \mathrm{E}(\mathbf{w}_t'(\mathbf{r}_{t+1} - \mathrm{E}(\mathbf{r}_{t+1})) = \mathrm{E}((\mathbf{w}_t - \mathrm{E}(\mathbf{w}_t))'\,\mathbf{r}_{t+1}). \tag{2}$$

Thus, holdings-based measures can be computed by de-meaning the portfolio weights or the returns, or both. Versions of all three approaches appear in the literature. However, in place of the expected weight or return we typically find a benchmark weight or return. The variables are re-centered using something other than the mean, which means that the empirical measures are not the covariances.

Holdings-based measures are estimated as versions of (1/T) Σ_i Σ_t w^i_t r^i_{t+1}, where either the weights w^i_t and/or the returns r^i_{t+1} are measured net of a benchmark. We use the notation without specifying the benchmark to simplify our presentation. The actual measures we study are benchmark adjusted, as:

$$GT = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T}((\mathbf{w}_t - \mathbf{w}_{t-\tau})'\mathbf{r}_{t+1}). \tag{3a}$$

$$DGTW = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{w}_t'(\mathbf{r}_{t+1} - \mathbf{r}_{t+1}^{D(t)})), \tag{3b}$$

$$\mathrm{CWM} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{w}_t - \mathbf{w}_{bh,t})'(\mathbf{r}_{t+1} - \mathrm{E}(\mathbf{r}_{t+1}|\mathbf{Z}_t)), \tag{3c}$$

$$\mathrm{FM} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\,'\mathbf{r}_{t+1}\,(\mathrm{a}-\mathbf{b}'\mathbf{r}_{Bt+1}), \tag{3d}$$

Equation (3a) is the portfolio change measure of Grinblatt and Titman (1993). This is an example of benchmarking the weights. The weight of the fund τ periods before the current period proxies for the expected weight. In this measure a manager records performance when the current portfolio $\mathbf{w}_t'\mathbf{r}_{t+1}$, achieves higher average hypothetical returns than the past portfolio weights, $\mathbf{w}_{t-\tau}$, would earn on the same returns. It is important to sum the covariances across the stocks because an informed manager may overweight some stocks and thus underweight others, such that the predictive ability may not be seen on subsets of stocks.

Grinblatt and Titman (1993) discuss the choice of the lag, τ. If τ is too small, the past weights might still contain information about the future stock returns, leading to an underestimation of the information in the current weights. If τ is too large, the portfolio's risk might change between the two dates and the measure, because it involves no risk adjustment, might be biased. We adopt the same criteria as Grinblatt and Titman (1993). We use $\tau = 4$ with quarterly data and $\tau = 12$ with monthly data.

Equation (3b) is the Characteristic Selectivity measure of Daniel, Grinblatt, Titman and Wermers (DGTW, 1997). The benchmark return vector $\mathbf{r}_{t+1}^{D(t)} = [r_{t+1}^{D_1(t)}, \cdots, r_{t+1}^{D_N(t)}]'$ is matched to each stock's size, book/market ratio and past momentum. Intuitively, an informed stock picker's portfolio of stocks can beat its portfolio-weighted combination of the DGTW benchmark returns.[4] This is an example of re-centering the stock returns, in the sense that $E(\mathbf{r}_{t+1} - \mathbf{r}_{t+1}^{D(t)})$ should be zero. In the data, of course, the mean of the recentered returns are not identically zero, and the mean of this difference is the vector of alphas.

Equation (3c) is the Ferson and Khang (2002) Conditional Weight-based Measure (CWM). This is an example of re-centering both the stock returns and the portfolio weights. The benchmark weight vector, $\mathbf{w}_{bht}$, is the actual weight from τ periods ago, updated with a buy-and-hold strategy: $w^i_{bht} = w^i_{t-\tau} \Pi_{j=1,\ldots\tau} [R^i_{t-\tau+j} / \Sigma_i w^i_{t-\tau+j-1} R^i_{t-\tau+j})]$, where R^i_t is the gross (one plus the rate of) return

[4] The DGTW benchmark return for each stock is constructed as follows. First, stocks are ranked by firm size and divided into five size groups, with each group having the same number of stocks. Within each size group the stocks are ranked by their market-to-book values, and divided into five market-to-book groups. Finally, in each of the 25 groups, the stocks are sorted by their average returns during the past months (*t*-2 to *t*-12) before the current month *t*, and split into five groups according to their past average returns. This produces 125 stock groups, each containing the same number of stocks. The value-weighted returns of the stocks in each of the 125 groups become the DGTW benchmark returns. Each stock is assigned one of the 125 benchmarks based on the closest match to its size, book/market and past returns.

of stock *i* at the subscripted date. The stock returns are demeaned using $E(\mathbf{r}_{t+1}|Z_t)$, the conditional mean returns estimated using regressions of the returns on lagged instruments. Because the benchmarked returns are actually mean zero, the measure delivers the covariance, which is the TSA. We use the first four lagged instruments in Ferson and Khang (2002): a lagged market dividend yield, default spread, term spread and short term Treasury bill yield.

The assumption of the CWM is that an informed manager departs from a buy and hold strategy when she can predict returns better than if using the public information. A fund delivers performance in the CWM when the portfolio's hypothetical unexpected return, based on the public information, exceeds that of the buy-and-hold benchmark. The CWM records zero performance, giving managers no credit for the mechanical use of the public information in Z_t, consistent with semi-strong form efficient markets in the sense of Fama (1970).

Equation (3d) is the Ferson and Mo (FM, 2016) measure. This is a version of the SDF (unconditional) alpha, $\alpha_p = E(m_{t+1}\ r_{pt+1})$, where $r_{pt+1} = \mathbf{w}_t`\mathbf{r}_{t+1}$ is the fund's hypothetical portfolio excess returns and m_{t+1} is the stochastic discount factor. Ferson and Mo assume a linear factor model for the SDF:

$$m_{t+1} = (a\text{-}\mathbf{b}'\boldsymbol{r}_{Bt+1}), \tag{4}$$

where $\boldsymbol{r}_{Bt+1}$ is a vector of K benchmark excess returns. The FM measure replaces r^i_{t+1} with the risk-adjusted excess stock return, $r^i_{t+1}\ (a\text{-}\mathbf{b}'\boldsymbol{r}_{Bt+1})$. This is an example of re-centering the stock returns, as $E_t\{m_{t+1}\ r^i_{t+1}\}$ should be zero. In the data the mean of the re-centered returns is not identically zero and $E\{m_{t+1}\ r^i_{t+1}\}$ is the vector of alphas.

A fund delivers abnormal performance in the FM measure by over-weighting stocks with subsequently high risk-adjusted returns. Ferson and Mo (2016) consider different choices for the benchmark return vector, $\boldsymbol{r}_{Bt+1}$, including the Carhart (1997) four factor model that we use here. We estimate the parameters (a,**b**) as discussed below, by requiring the model to correctly price the K factors $\boldsymbol{r}_{Bt+1}$ and a Treasury bill return.

The holdings-based performance measures are hypothetical returns on a *before-cost* basis. They are not the returns to investors, who must bear funds' turnover-related trading costs, expense ratios and other investor costs. As discussed in Grinblatt and Titman (1989b) we are more likely to find evidence of skill in the before-cost returns than in the after-cost returns of funds.

2.2 TSA and AAR in holdings-based performance measures

The classical holdings-based measures are versions of $(1/T)\ \Sigma_i\ \Sigma_t\ w^i_t\ r^i_{t+1}$, where the security level variables r^i_{t+1} and w^i_t are recentered or benchmark adjusted, but not necessarily mean zero. The classical measures reflect both TSA and AAR. The expression of these components depends on whether a measure benchmark adjusts the returns or the weights.

In some cases only the returns are benchmark adjusted (DGTW, FM). Let the stock-specific benchmark be $\mathbf{r}_B$ (with betas and traded factors, **f**, then $\mathbf{r}_B = \beta\ \mathbf{f}$). Let $\boldsymbol{\alpha} \equiv E[\mathbf{r}- \mathbf{r}_B]$,. The expected performance measure is $E\{\mathbf{w}'[\mathbf{r}- \mathbf{r}_B]\} = Cov\{\mathbf{w}'; [\mathbf{r}- \mathbf{r}_B]\} + E(\mathbf{w})'\boldsymbol{\alpha}$. The first term is the TSA and the second is the AAR.

In some cases only the portfolio weights are adjusted (Cornell (1979), GT, Kacperczyk et. al, 2014).[5] The expected performance is $E\{[\mathbf{w}- \mathbf{w}_B]'\mathbf{r}\} = Cov\{[\mathbf{w}- \mathbf{w}_B]'; \mathbf{r}\} + E[\mathbf{w}- \mathbf{w}_B]'\ E(\mathbf{r})$**,** where

[5] Kothari and Warner (2001) advocate using trades, or changes in holdings. They use an event study approach, in which trade dummies appear in a panel regression for returns. Kacperczyk et al. (2014) benchmark the weights using market value weights: $\sum_i (w_{it}-w_{mt})\ r_{it+1}$. With a factor model for the returns, $r_{it+1} = \alpha_i + \boldsymbol{\beta}_i{}'\ \mathbf{r}_{Bt+1} + u_{it+1}$, the measure decomposes into a timing and a stock picking term: $\sum_i (w_{it}-w_{mt})(\beta_i{}'$

$\mathbf{w}_B$ is the vector of benchmark weights. If E[$\mathbf{w}$- $\mathbf{w}_B$] is not a vector of zeros, the measures have both TSA and AAR components. Both are measured relative to the benchmark weights. The AAR measures the return attributed to the average deviation from the benchmark weights. The average deviation from benchmark weights could be a policy decision of the fund. In the CWM both returns and weights are benchmarked, and the AAR is zero.

3. Holdings-based Measures are Panel Regressions

3.1. A Basic Panel Regression

Many previous studies have employed panel regressions in mutual fund research. Typically, however, the left-hand side variables are measured at the fund level. For example, Ippolito (1992) and Sirri and Tufano (1998) regress the new money flows of mutual funds on past measures of fund performance. Other studies put measures of fund performance on the left hand side to investigate its relation to things like fund size (Ferreira, Keswani, Miguel and Ramos, 2013), fund and industry size (Pastor, Stambaugh and Taylor, (2015), Zhu (2018), Magkotsios, 2018) or other variables.

In the panel regression view of holdings-based performance measures the returns of the stocks held by a fund are on the left hand side and a fund's holdings of the stocks are on the right hand side. There is a separate panel regression for each fund and a fund-specific performance measure. Starting with the simplest case, assume that a fund's portfolio contains *N* stocks that exist for *T* periods. The panel regression is:

$\mathbf{r}_{Bt+1}$) + $\sum_i$ (w_{it}-w_{mt})(α_i +u_{it+1}). They find that funds vary their emphasis on the two terms over the business cycle.

$$r_{t+1}^{i} = \beta w_t^i + \varepsilon_{t+1}^{i}., \ i=1,\ldots,N; \ t=1,\ldots T. \qquad (7)$$

The slope coefficient β captures the ability of the fund's weights to predict future excess stock returns. The pooled OLS slope coefficent estimator in the regression (7) is:

$$\hat{\beta} = \Sigma_i \, (1/T)\Sigma_t \, (r^i_{t+1} \, w^i_t) \, / \, \Sigma_i \, (1/T)\Sigma_t \, (w^i_t \, w^i_{t'}) \qquad (8)$$

For comparison, consider the estimator of fund performance $DGTW = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{w}_t{}'(\mathbf{r}_{t+1} - \mathbf{r}_{t+1}^{D(t)}))$, which is a version of the numerator in Equation (8), where the benchmark adjusted returns replace the returns r^i_{t+1}. The numerator of (8), Σ_i (1/T)Σ_t w^i_t r^i_{t+1}, also includes the other holdings-based measures as special cases.[6] Using the weight change, w_{it} - $w_{it-\tau}$, in place of w^i_t we obtain the GT measure. Replacing w_t with (w_t - w_{bt}) and $r^i_{t+1}$with [r^i_{t+1}- E(r^i_{t+1}|Z_t)], we obtain the CWM. Finally, replacing r^i_t with r^i_t (a-$\mathbf{b}$'$\mathbf{r}_{Bt+1}$) we obtain the FM measure. Tests of the hypothesis of zero performance ask if the coefficient β is zero.

3.2 Fixed Effects

Equation (7) is unrealistic for stock returns, because the mean values of the stock returns vary in the cross-section. The expected stock returns and the portfolio weights are both positive. Imposing a zero intercept like in (7) would produce a positive slope coefficient estimate even if there was no relation between between the returns and weights. Introducing stock fixed effects allows each

[6] Note that both the numerator and denominator of (8) can be divided by N, but because the weights sum to 1.0, it is better not to. Assume that the expected value of the stock return r_i is bounded on [r_L,r_u]. Because funds rarely sell short, the portfolio weights {w_i} are all between 0 and 1 and they sum to 1.0. Therefore, E{$\Sigma_i w_i r_i$} is bounded on [r_L,r_u], so E{(1/N) $\Sigma_i w_i r_i$}$\leq$ r_u/N -> 0. The measures approach zero for large portfolios if divided by N.

stock to have its own mean and the regression slope captures the covariance of returns and weights. The panel regression model with stock fixed effects is:

$$r_{t+1}^{i} = a^{i} + \beta w_{t}^{i} + \varepsilon_{t+1}^{i}. \qquad (9)$$

The model is estimated by introducing *N* stock dummy variables, which take the value of one (for each date) if the return belongs to stock *i* and zero otherwise. The coefficient of the dummy variable for stock *i* is a^{i}, the fixed effect of stock *i*. Under the null hypothesis ($\beta = 0$), the fixed effect is the expected excess return of the stock.

The classical performance measures reflect both time-series and cross-sectional variation. Fixed effects separate these two dimensions. With stock fixed effects in the model, the Frisch–Waugh (1933) theorem shows that the OLS slope coefficient estimator of the regression (9) in the same as that obtained by subtracting the time-series means from each of the variables and running the regression on the demeaned variables with no intercept. Thus, the numerator of the OLS slope estimator of (9) estimates the TSA:

$$\hat{\beta}_{w,num} = \Sigma_{i}\,(1/T)\,\Sigma_{t}\,[r^{i}_{t+1} - (1/T)\Sigma_{t}\,r^{i}_{t+1}][w^{i}_{t} - (1/T)\Sigma_{t}\,w^{i}_{t}] \qquad (10)$$

$$= \mathrm{Cov}(\mathbf{w}_{t}';\mathbf{r}_{t+1}).$$

It follows from (10) that:

$$\Sigma_{i}\,(1/T)\,\Sigma_{t}\,(r^{i}_{t+1}\,w^{i}_{t}) = \hat{\beta}_{w,num} + \Sigma_{i}\,[(1/T)\Sigma_{t}\,r^{i}_{t+1}]\,[(1/T)\Sigma_{t}\,w^{i}_{t}]. \qquad (11)$$

Thus, the numerator of the classical performance measure on the left-hand side of (11) is equal to the numerator of the estimator with stock fixed effects, plus an AAR term. If r^i_{t+1} is replaced with a factor-adjusted return (in DGTW and FM) whose mean is α_i, the second term of (11) is an estimate of $\boldsymbol{\alpha}$'E($\boldsymbol{w}$). Comparing the panel regressions with and without stock fixed effects, we decompose the classical measures into the TSA and the AAR.

The Internet Appendix discusses models with time fixed effects, both time and stock fixed effects, and a model with a common intercept. Because the portfolio weights sum to one, a regression with time dummies and a regression with a common intercept have the same effects on the slope coefficient. The numerator of the slope with time dummies differs from that of the basic panel regression with no dummies by the mean of an equally weighted portfolio return. Because the equally-weighted portfolio return should be similar across funds, the regressions with time dummies or a common intercept provide little additional insight about the structure of fund performance.[7]

3.3 Scaling the Panel Regressions

In order to test the null hypothesis of zero performance the panel regression can be used directly, as the slope coefficients are zero under the null hypothesis of no performance. Panel regressions are easily rescaled to recover the numerator of the slope coefficient, which may be more economically meaningful. For example, numerator in the FM panel regression is a certainty equivalent excess return for an agent with the SDF used in the model.

[7] It might be natural to consider a single cross-sectional regression with an intercept for returns on lagged weights for each time period, as in Fama and MacBeth (1973). This delivers an OLS slope coefficient for period *t* whose numerator is equal to [r_{pt+1} - $r_{EW,t+1}$], where $r_{pt+1} = \mathbf{w}_t'\mathbf{r}_{t+1}$ and $r_{EW,t+1}$ is the equally weighted portfolio return. The time-series average of the Fama-MacBeth slopes is a weighted average difference between the fund's and the equally-weighted portfolio's returns.

To illustrate the scaling let Let u_r = Vec(r_{it+1}-E(r_i)), u_w = Vec(w_{it}-E(w_i)), u_x = Vec(x_{it}-E(x_i)). These have the same probability limits as the residuals from regressing r,w and x on stock dummies, according to the Frisch-Waugh (1933) theorem. We scale the TSA by regressing (u_w'u_w/T) u_r on u_w. The OLS slope is $(u_w'u_w/T)^{-1}$ (u_w'u_r/T) (u_w'u_w/T), which converges in probability to Cov(w_t'; r_{t+1}). With the Hjalmarsson (2010) and DiffIV methods described below, we use instrumental variables instead of OLS. Giles (1984) shows that the Frisch Waugh (1933) theorem holds for instrumental variables estimation. We carry these scaling conventions to the simulations to evaluate the estimators' statistical properties.

3.4 Standard Errors

Ferson and Khang (2002) use time-series GMM-derived asymptotic standard errors for their CWM, while this paper concentrates on a panel regression interpretation. Consider the pooled OLS estimate of the panel regression (7), where $\hat{\beta} - \beta = [(1/T)\, \Sigma_i\, \Sigma_t\, (w^i_t\, w^i_t{}')]^{-1}\, (1/T)\, \Sigma_i\, \Sigma_t\, (\varepsilon^i_{t+1}\, w^i_t)$ and

$E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\} \approx [(1/T)\, \Sigma_i\, \Sigma_t\, (w^i_t\, w^i_t{}')]^{-1}\, E\{(1/T)\, \Sigma_i\, \Sigma_t\, (\varepsilon^i_{t+1}\, w^i_t)\}^2\, [(1/T)\, \Sigma_i\, \Sigma_t\, (w^i_t\, w^i_t{}')]^{-1}$.

Using the conventional assumption that $[(1/T)\, \Sigma_i\, \Sigma_t\, (w^i_t\, w^i_t{}')]^{-1}$ converges in probability to a constant, $(\sigma_x{}^2)^{-1}$, different ways of computing the central term, $E\{(1/T)\, \Sigma_i\, \Sigma_t\, (\varepsilon^i_{t+1}\, w^i_t)\}^2$, correspond to different methods of clustering in the panel. Petersen (2009) focusses on panel standard errors under various models for the error terms. For example, clustering by time assumes no time series dependence, so $E\{(\varepsilon^i_{t+1}w^i_t)\,(\varepsilon^i_{\tau+1}w^i_\tau)\}=0(t\neq\tau)$ and $E\{(1/T)\Sigma_i\Sigma_t\,(\varepsilon^i_{t+1}\, w^i_t)\}^2 = E\{(1/T)\, \Sigma_t(\Sigma_i\, \varepsilon^i_{t+1}\, w^i_t)^2\}$. This term is estimated with the fitted regression residuals. Keeping the stocks within the time cluster allows for dependence across the stocks. Clustering by stock we assume no cross-sectional dependence: $E\{\, (\varepsilon^i_{t+1}\, w^i_t)\,(\varepsilon^j_{t+1}\, w^j_t)\}=0\ (i\neq j)$ and $E\{(1/T)\, \Sigma_i\, \Sigma_t\,(\varepsilon^i_{t+1}\, w^i_t)\}^2 = E\{(1/T)\, \Sigma_i\, (\Sigma_t\, \varepsilon^i_{t+1}\, w^i_t)^2\}$. Keeping the times within the cluster for a stock allows for dependence over time.

GT and DGTW use time-series methods to compute standard errors. For example, DGTW define $CS_t = \Sigma_i (r^i_{t+1} w^i_t)$, where the returns are benchmark adjusted as explained above, but we keep the simpler notation r^i_{t+1} here. They define CS as the time-series average $(1/T) \Sigma_t CS_t$. Plugging in the regression we express $CS_t = \Sigma_i ([\beta w^i_t + \varepsilon^i_{t+1}] w^i_t') = \beta \Sigma_i (w^i_t w^i_t') + \Sigma_i w^i_t \varepsilon^i_{t+1}$ and $CS = \beta (1/T) \Sigma_t \Sigma_i (w^i_t w^i_t') + (1/T) \Sigma_t \Sigma_i w^i_t \varepsilon^i_{t+1}$. The first term is the "true" scaled slope coefficient and the second term is the sampling error in its estimate. DGTW use time series methods to estimate the variance of the sampling error. The calculation $(1/T)\Sigma_t CS_t^2 - [(1/T)\Sigma_t CS_t]^2$ delivers the sample variance and corresponds to clustering by time in the pooled panel regression, because terms like $(\Sigma_i w^i_t \varepsilon^i_{t+1})^2$ are kept within the time cluster.

Since monthly stock returns are cross-sectionally correlated at a point in time, but have little time-series correlation, we cluster the standard errors by time, following DGTW. When using recursively demeaned or differenced estimation as described below, we include Newey and West (1987) autocovariance terms to accommodate the serial dependence induced by overlapping data. In this case $E\{\Sigma_i \Sigma_t (\varepsilon^i_{t+1} w^i_t)\}^2 = E\{\Sigma_t (\Sigma_i \varepsilon^i_{t+1} w^i_t)^2\} + E\{\Sigma_\tau \Sigma_{t\neq\tau} K(|t-\tau|) \Sigma_i (\varepsilon^i_{t+1} w^i_t)(\varepsilon^i_{\tau+1} w^i_\tau)$, where $K(|t-\tau|)$ is the Bartlett kernel. We use 30 lags to control the autocorrelation.

3.5 Lagged Stochastic Regressor Bias

Predictive panel regressions are affected by a bias due to lagged stochastic regressors, known as the "Stambaugh bias" in panels, related to the persistence of the portfolio weights and to the correlation of their future values with the innovations in the stock returns. The bias is examined in time-series regressions for stock market returns on dividend yields by Stambaugh (1999), Pastor and Stambaugh (2009), Amihud and Hurvich (2004) and Amihud *et.al.* (2008, 2010) and by Zhu

(2018) in panel regressions of fund performance on fund size. We examine several methods to address the bias. These include a parametric bias correction (Hjalmarsson, 2008), a recursively demeaned instrumental variables approach, (Hjalmarsson, 2010) and a differenced instrumental variables approach (DiffIV, Anderson and Hsiao (1981), Wang, 2015).

To capture the fact that the fund's portfolio weights are highly autocorrelated, the parametric bias correction assumes that they follow a first-order autoregression:

$$w_{t+1}^{i} = \gamma^{i} + \rho w_{t}^{i} + v_{t+1}^{i}. \tag{12}$$

We later discuss alternative models for the portfolio weights that depart from the AR(1) assumption of (12).

The lagged stochastic regressor bias in the panel regression with fixed effects may be understood as follows. Consider the numerator of the Within estimator, written without loss of generality with only the weights demeaned:

$$\hat{\beta}_{w,num} = (1/T)\Sigma_{i}\, \Sigma_{t}\ \ r^{i}_{t+1}\, (w^{i}_{t} - (1/T)\Sigma_{t}\, w^{i}_{t}), \tag{13}$$

Substituting from Equation (9), on the assumption that the shocks to the stock returns, ε^{i}_{t+1}, are uncorrelated with lagged weights under the null hypothesis of no ability, yields:

$$E(\hat{\beta}_{w,num} - \beta_{num}) = -E[(1/T)\Sigma_{i}\, \Sigma_{t}\, \varepsilon^{i}_{t+1}\, (1/T)\Sigma_{\tau\geq t+1}\, w^{i}_{\tau}]. \tag{14}$$

Equation (14) shows that a bias arises if ε^{i}_{t+1} is correlated with $\Sigma_{\tau\geq t+1}\, w^{i}_{\tau}$. For example, we would expect a positive correlation between ε^i_{t+1} and future weights under a buy-and-hold portfolio strategy, as a positive return shock increases the portfolio weight. We expect a negative bias in the slope estimate. The larger is the serial correlation in the weights, the larger is the expected bias as the return shocks at time *t+1* accumulate in the future weights, w^i_{τ}, $\tau \geq t+1$. Hjalmarsson (2010) shows that the bias is a second order bias. Specifically, the estimator is consistent as both N and T become large. However, when T is finite and N becomes large, the bias remains because the number of fixed effects to be estimated expands with N.

3.6. Additional Model Parameters

For some of the classical performance measures parameters for risk adjustment must be estimated in addition to the panel regression coefficients. For example, in the FM measure the parameters (a,b) in the SDF $m_{t+1} = a - \boldsymbol{b}'r_{Bt+1}$ must be estimated. Simultaneous estimation of these parameters is usually preferable. Consider a pooled generalized method of moments (GMM, Hansen, 1982) system with $g=(1/NT)\Sigma_i\Sigma_t\, g_{it}$, where g_{it} is a K+2 vector when there are K traded excess return factors in r_B and $\boldsymbol{b}$ is a K-vector of parameters. The first K elements of g_{it} are given by $r_{Bt+1}(a-\boldsymbol{b}'r_{Bt+1})$ and the K+1st element is $R_{Ft}(a-\boldsymbol{b}'r_{Bt+1}) - 1$, where R_{Ft} is the gross risk-free rate. The K+2nd element of g_{it} is $[r^i_{t+1}(a-\boldsymbol{b}'r_{Bt+1}) - \beta w^i_t]w^i_t$ when Equation (7) is the panel regression, but we do not divide by N for this term because the weights sum to 1.0. With a four factor model for $\mathbf{r}_B$ there are six parameters and six moment conditions and the GMM system is exactly identified. (The other models also produce exactly identified systems.) The GMM parameter estimates may be found by setting g=0. Simultaneous estimation of $(a,\boldsymbol{b},\beta)$ produces the same point estimates as does a two-step approach, where the parameters $(a,\boldsymbol{b})$ are found in the first step, using the first K+1 elements of g, and the parameter β is found by estimating the panel regression using OLS, substituting in the first step parameter estimates for risk adjustment. We take this two-step approach.

The GMM standard errors for the two step and simultaneous approaches are not the same, as the standard errors for the OLS panel regression in the second step do not reflect the estimation error in the parameters (a,***b***). We compare the standard errors for the simultaneous estimation with those of the two-step calculation and find that they are smaller. Thus, the t-ratios with the standard errors from the two-step estimation will be smaller, a conservative bias. We explore the magnitudes of this bias with simulation and find it to be small.[8]

4. Data

We use quarterly holdings data for mutual funds from Thomson-Reuters for 1980 to 2012 and we apply several standard screens. We exclude data before 1984 in most of our analyses. Fama and French (2010) show that there is a selection bias in mutual fund data before 1984. Since most of the research papers using holdings data focus on US equity funds, we exclude other types of mutual funds. Evans (2010) discusses an incubation bias in fund performance measures, and following his suggestions, we exclude observations before the reported date of fund organization or when a fund first has total net asset value (TNA) of less than 15 million dollars. There are in total 3596 equity funds in this sample. We also collect the holdings for a sample of 201 index funds for 1994-2012, identified by either the CRSP index fund flag or by searching for the string "index" in the funds' name.

We use monthly prices and returns of individual stocks from the CRSP monthly stock file,

[8] Partition the joint system as $g=(g_1',g_2')'$ and consider the form of the asymptotic covariance when the system is exactly identified. Its estimate in the full system is $\text{ACov}(\beta) = \text{Cov}(g_2)/\delta^2 - \mu' A^{-1}\text{Cov}(g_1,g_2)\ )/\delta^2$, where $\delta = (1/T)\Sigma_i\Sigma_t w^i_t{}^2$, $\mu = -\ (1/T)\Sigma_t (\Sigma_i w^i_t r^i_{t+1})\ r_{Bt+1}$, and $A= (1/NT)\Sigma_i\Sigma_t (r_{Bt+1}\ r_{Bt+1'})$. The term $\mu' A^{-1}$ is the regression coefficient for the time-series regression of the funds' hypothetical portfolio return on the benchmark excess returns. The two-step calculation uses a subset of the moment condition and delivers $\text{ACov}(\beta) = \text{Cov}(g_2)/\delta^2$. If the subset moment conditions are uncorrelated, the two-step and joint estimation deliver the same standard errors.

including the delisting returns when firms leave the market. The sample contains stocks with at least one month of return data from 1970 to 2012. The DGTW benchmark returns are collected from Russ Wermers' website, and are then combined with the monthly stock file. We select only stocks with non-missing values of the DGTW benchmark returns. According to Wermers (2006), the stocks selected for the DGTW benchmarks have at least two years of data on book values, returns and market capitalization. For some of our analyses we use daily stock returns and mutual fund returns from CRSP, available starting in 1999 for the mutual funds. We also use daily data on interest rates from the Federal Reserve Data base.

We use holdings and stock prices to construct portfolio quarterly and monthly weights for the mutual funds. We first describe the quarterly weights. Let the holdings of the stocks (measured as the number of shares held) and stock prices at the t'th quarter be $\mathbf{h}_t$ and $\mathbf{p}_t$, where $\mathbf{h}_t$ = [$h_{1t},\ldots,h_{Nt}$]' and $\mathbf{p}_t$ = [$p_{1t},\ldots,p_{Nt}$]'. The weights of the stocks in the fund portfolio are $w_t^i = \frac{h_{it}\, p_{it}}{\boldsymbol{h}_t{}'\, \boldsymbol{p}_t}$. The universe of stocks for a fund is the set of stocks ever held by a fund. Following GT and DGTW we not not measure holdings of bonds or cash, normalizing the weights in stocks to sum to 1.0. (In the Internet Appendix we incorporate cash holdings earnging at a risk-free rate and find similar results.) Since the weights for never-held stocks in a given quarter are zero, including any additional stocks in the universe would not change the performance measures.[9]

We construct monthly weights, assuming that for the months between consecutive reporting dates the funds keep holding the same number of shares, except that a fund reinvests any dividends in the same stock. Thus, the monthly portfolio weights are not interpolated, but updated with the

[9] The buy-and-hold drift measures do depend on the universe of stocks. We conduct some experiments where we expand the universe for each fund to include 1,000 stocks ever held by any fund in the sample. When the buy-and-hold drift strategy buys a new stock because some other stock leaves the sample, it chooses randomly from the 1,000. The results are similar.

actual return data, on the assumption of a buy-and-hold strategy between the quarters. Filling out quarterly weights to monthly data like this follows DGTW (1997), Kacperczyk et al. (2005), Busse and Tong (2012) and others. The flow performance analysis below uses monthly flows, and for comparability with earlier studies and we use the monthly weights to form monthly performance measures. Elton, Gruber and Blake (2011) find that actual monthly weight data afford greater precision than quarterly, although the fund performance measures (alphas) have a correlation greater than 99%. This impact of this convention is that the quarterly measure is roughly the compounded value of the monthly measures.[10] We find similar results using either the quarterly or the monthly measures, as shown in an Internet Appendix.

4.1 Summary Statistics

Summary statistics for the data are shown in Table 1. The sample period is from 1984 to 2012, and the number of funds is 3596. In panel A, for each fund in the sample we compute the time-series average of its total net assets (TNA) in millions of dollars, the average number of stocks held, the Return Gap from Kacperczyk, Sialm and Zheng (2006) and the Active Weight as in Doshi et al. (2015). We compute the sample standard deviations of the monthly reported fund returns, σ, and their R-squares in Carhart (1997) four-factor model regressions. We also compute the sample autocorrelations of the funds' portfolio weights, ρ_j, averaged across the holdings, at various lags j, j=1,…5. The means, std, maximum and minimum are taken across the funds in the sample. Finally, for each fund we compute the correlation between the errors of the portfolio weight autoregression and the panel regression of future stock returns on the weights, averaged across the holdings. This

[10] Let the weights of fund in securities *i* at the beginning of a quarter be W_{0i}, and the gross returns of the securities over the next three months be R_{1i}, R_{2i}, and R_{3i}. The quarterly measure is $\Sigma_i W_{0i} (R_{1i} R_{2i} R_{3i})$ -1. The weights after one month are $W_{1i} = W_{0i} (R_{1i} /\Sigma_i W_{0i}R_{1i})$ and the weights at month two are $W_{2i} = W_{1i} (R_{2i} / \Sigma_i W_{1i} R_{2i})$. Thus, the quarterly measure is $[\Sigma_i W_{0i} R_{1i}] [\Sigma_i W_{1i} R_{2i}] [\Sigma_i W_{2i} R_{3i}]$ -1.

average correlation, denoted as Error Corr in the table, serves as Cov(ε_i,v_i) in the bias-adjusted estimator of Hjalmarsson (2008). The maximum and minimum values are 0.24 and -0.19, respectively, which suggests that the Stambaugh bias can change signs for different funds.

The average fund has total assets of $684 million and holds 114 stocks. The average first order autocorrelation of the weights at the stock level is 0.90, and 0.94 at the fund level. The weights are persistent time-series, but as they are bounded on [0,1] they are likely stationary. We report simulations below that allow for trends in the weights and long strings of zero weights, like we find for some stocks. The first order autocorrelaiton is below the level where the simulation evidence in Ferson, Sarkissian and Simin (2003) indicates that spurious regression becomes a concern. However, the autocorrelation is high enough to make the lagged stochastic regressor bias a concern.

For some of our simulations we assume a first order autoregressive process (AR(1)) for the weights. As a reality check we estimate the coefficients ρ_j in the following monthly regressions: $w_{t+1}^i = \gamma^i + \rho_j w_{t+1-j}^i + v_{t+1}^i$, where j runs from 1 to 5. If the weights follow an $AR(1)$ process, we should observe that $\rho_j = \rho_1^j$. Panel A of Table 1 suggests this is a good approximation. For example, at the mean values, ρ_1^2 = 0.815 and ρ_2 = 0.813, while ρ_1^5 = 0.599 and ρ_5 = 0.586. We also compute the autocorrelations for the quarterly weights data and report them in the internet appendix. In this case ρ_1 = 0.83, ρ_2 = 0.69 and ρ_4 = 0.51, with ρ_1^2 = 0.69 and ρ_1^4 = 0.47.

4.2 Active and Passive Components in AAR

The AAR component of performance is the performance of the average, or policy weights of a fund. If a fund adopts a strategy of actively rebalancing to fixed policy weights, its portfolio weights would vary little over time and all of its performance would be AAR. However, AAR also arises from passive strategies, such and buy-and-hold, where the portfolio weights drift. The portfolio weights of a buy-and-hold strategy drift over time towards the higher return and thus higher alpha stocks, which increases AAR. We call this effect the *buy-and-hold drift* and find that it is a substantial piece of some classical performance measures. Since it can be useful to separate active from passive components of performance, we extract the buy and hold drift component from the AAR.

Table 2 illustrates the AAR from two buy-and-hold strategies and shows that it can be substantial. Panel A starts in 1980, holding the stocks held by any mutual fund in our sample, with weights equal to the aggregate total net asset values. Panel B starts with the stocks in the Standard and Poors 500 (SP500) index, held with market capitalization weights. The share holdings are fixed until 2012, except that dividends are reinvested in the dividend-paying stocks. The weights thus evolve as a buy-and-hold strategy. If a stock leaves the index, the strategy sells the stock and puts the money in the entering stock(s). If a stock delists, the value after the delisting return is divided among the remaining stocks.

Table 2 shows that the weight in the 25 highest DGTW alpha stocks starts in 1980 at 2.71% and finishes in 2012 at 8.02%, after buying and holding. The average weights are higher for the high-alpha stocks. The differences between the high and low alpha average weights are 1.19% to 2.74% across the models. The bottom rows of each panel show the AARs of the buy-and-hold strategies, which vary from 0.25% to 0.93% per year across models[11] Table 2 establishes that a

[11] It may seem puzzling that a buy-and-hold strategy can have positive AAR. The portfolio-weighted average of stocks' alphas in a factor model regression must be zero when the weights correspond to those of the market or factor mimicking portfolios in the model, on the assumption that the weights are fixed or have no information about the future factor returns or factor model residuals. This suggests that the AARs of the market and factor mimicking portfolios should be zero or close to zero. A buy and hold strategy might be similar to the market portfolio. However, not all of the stocks in the market portfolio are held by the funds in our sample. We find 6961 stocks on CRSP for which holdings are never reported by any mutual fund during 1980-2012 (21,217 stocks are held at least

passive buy-and-hold strategy can generate AAR. Simulations reported in the Internet Appendix show that a stylized momentum strategy and a buy-and-hold strategy, where the weights depart from the AR(1) assumption and are calibrated to more realistically mimick a mutual fund, also produce positive AAR.

In the mutual funds we decompose the AAR in the classical performance measures to isolate the passive buy-and-hold drift component. Let $\mathbf{w}_{BH}$ be the weights of a buy-and-hold strategy, starting with the initial holdings of a fund. The buy-and-hold weights evolve passively over time, and are adjusted to keep holdings below 5% as described in the Internet Appendix. A fund's weights are $\mathbf{w} = \mathbf{w}_{BH} + (\mathbf{w} - \mathbf{w}_{BH})$ so that $E(\mathbf{w})'\boldsymbol{\alpha} = E(\mathbf{w}_{BH})'\boldsymbol{\alpha} + E(\mathbf{w}\text{-}\mathbf{w}_{BH})'\boldsymbol{\alpha}$. The first term is the passive buy-and-hold drift and the second term is the active AAR net of the buy-and-hold drift.[12] When the performance measure benchmarks only the weights with benchmark $\mathbf{w}_B$ we decompose the AAR as $E(\mathbf{w}\text{-}\mathbf{w}_B)'E(\mathbf{r}) = E(\mathbf{w}\text{-}\mathbf{w}_{BH})'E(\mathbf{r}) + E(\mathbf{w}_{BH}\text{-}\mathbf{w}_B)'E(\mathbf{r})$. The first term is the AAR in excess of buy-and-hold drift and the second term is the buy-and-hold drift.

5. Simulation Results

We evaluate various methods for correcting the lagged stochastic regressor bias. Two methods perform well in panels of the sizes in this paper, illustrated in Table 3. Our first approach to bias correction is a parametric approach from Hjalmarsson (2008) which is an extension of the approach of Stambaugh (1999) to panels. The second is a recursively demeaned approach from Hjalmarsson

once). The alphas of these never-held stocks are typically negative. For example the value-weighted DGTW alpha of these stocks is -1.89% per year. The equally weighted alphas are -1.55% under DGTW and -3.81% under the FM measure. Skripnik (2019) provides a detailed analysis of mutual funds' stock screening decisions.

[12] The average alpha of the initial stock selection is removed when the AAR net of buy-and-hold drift examined, resulting in $Cov_i(\alpha_i, E(w_i)\text{-}E(w_{iBH}))$. Thus, in the net AAR we measure the cross sectional covariance between the average or policy weights and the stocks' alphas, net of that for the passive buy-and-hold strategy.

(Haj10, 2010). The third is a differenced instrumental variables approach (DiffIV) similar to Anderson and Hsiao (1981) and Wang (2015).

5.1 Bias Corrections

We bootstrap versions of the predictive system (9) and (12) for various "true" values of the panel regression slope, β. We estimate the pooled panel model for β using the DiffIV method and sort funds based on the estimated slopes. We select five funds near the 5%, 10%, 50%, 90% and 95% cutoff values, and use the estimated β values and the holdings data of the five selected funds to calibrate the simulations. The data generating process features heterogeneous fixed effects across the stocks. The parameter ρ for the simulated weights is also allowed to differ across the stocks. For a given value of the slope β, the intercept for each stock is chosen to match the sample average benchmark adjusted returns. For a given value of ρ the intercepts in the weight's VAR equations are fixed to match the average values of the actual weights for each stock. We boostrap the vector of returns and weight residuals from equations (9) and (12) together for all of the stocks, selecting months randomly from the data with replacement. We build up the simulated weight and return series recursively, using the calibrated values of β, ρ and the intercepts.

The simulations preserve the serial dependence in the weights, dependence across the stock returns, and importantly, thedependence between the returns and weight innovations. The number of months for each fund in the simulations is the same as the number of months where the fund exists in the real data. When a stock held by the fund has a missing return, the weight is set to zero.

Table 3 presents the results. The true values of the slopes are shown in the first column and the estimated values, averaged across 1000 simulation trials, are shown in the other columns. The regressions are scaled to deliver the numerator of the slope coefficient, measured as an excess return per quarter. For comparison purposes we include results for the classical OLS estimator of Equation (7) with no fixed effect, OLS estimation with a common intercept, the classical Within estimator (OLS with stock dummies) and a first difference estimator (Diff). These comparison estimators have large biases as expected. The largest bias in the Diff IV estimator across the five experiments is only 0.03% per quarter. The winner is Haj10, where the largest bias is only 0.01% per quarter.

We use the difference between panel regression slopes with versus without fixed effects to estimate AAR, where the model with fixed effects is estimated using Haj10. Additional simulations in Panel B of Table 3 evaluate the accuracy to measure AAR.[13] The estimated AAR for the various performance measures is averaged across 1,000 simulation trials. The evidence supports the validity of our approach to estimating AAR. The maximum difference between the true and expected estimated AAR is 0.08% per quarter and the average difference is less than 0.04%.

The Internet Appendix presents simulations of hypothetical mutual fund strategies, including a style strategy based on momentum and a buy-and-hold strategy. These simulations do not rely on the AR(1) assumption for the portfolio weights, and they have realistic sequences of zero and nonzero weights. There is no TSA in these strategies, and the approach correctly infers this. Both strategies produce significant positive AAR estimates. The distributions of the t-ratios for the TSA estimates are well approximated by a unit normal distribution.

[13] The "true" AAR in the data generating process of (9) and (12) differs across the measures. This is given by $E(\mathbf{w})'E(\mathbf{r}) = \Sigma_i a^i E(w_i) + \beta \Sigma_i E((w_i)^2)$, where the "true" parameters for this calculation come from an initial simulation that assumes all the stocks for the chosen fund exist each month. The second term is small and goes to zero as the number of stocks, N, gets large. For example, it is equal to β/N for an equally-weighted portfolio.

6. Empirical Results

6.1 Cross-sectional Variation in Fund Performance

We examine the classical holdings-based performance measures in cross-sections of individual funds. Our interest is to understand how and why the measured performance differs across funds; in particular to assess the influence of TSA and AAR.

We sort funds into five quintiles on the basis of each classical performance measure, estimated using the full sample of data for each fund. The averages for each quintile are shown as the left-hand columns of numbers in Table 4. We also compute the Haj10 estimates of TSA for each fund. The difference between the original measure and the TSA is the AAR. In some cases the AAR is further decomposed into the buy-and-hold drift and the AAR in excess of buy and hold. The averages of the various components for each original-measure quintile are shown in the remaining columns of the table.

The bottom rows of each panel of Table 4 are the differences between the average performance measures for the top and bottom quintiles (HML), summarizing the cross-fund spreads. Sorts on the classical measures produce large HMLs of 7.69% per year for GT, 6.61% for DGTW, 6.05% for the CWM and 27.72% for FM.

The drivers of the spreads are different for the different measures. For GT the spread in TSA is large, with an HML of 6.48%, so the GT measure differences reflect mainly TSA differences. The GT measure and its TSA show positive skewness. The very best funds under GT are likely to have very high TSA. The AAR component of GT shows a small spread and we do not break it down further in Table 4.

Unlike the GT measure, a small portion of the cross-sectional variation in the DGTW measure is driven by TSA. The AAR is the main driver. Its HML is 5.21%, representing more than 80% of the total spread in DGTW. The sample correlation across individual funds between their DGTW measures and the AAR component is 0.69. Sorting funds on their DGTW measures is a lot like sorting them on their AAR. Furthermore, more than 2/3 of the AAR spread is associated with the buy-and-hold drift. Thus, differences among funds' DGTW performance measures, to a large extent, capture differences in the passive buy-and-hold drifts.[14]

Panel C of Table 4 presents an unbiased version of the conditional weight measure, CWM, estimated using the Haj10 approach. The CWM is 100% TSA. We use the DGTW benchmarks to break its TSA into timing, Cov($\mathbf{w}$-$\mathbf{w}_{BH}$'; $\mathbf{r}_B$ - E($\mathbf{r}_B$|Z)) and selection, Cov($\mathbf{w}$-$\mathbf{w}_{BH}$'; $\mathbf{u}$-E($\mathbf{u}$|Z)). The buy-and-hold benchmark weights $\mathbf{w}_{BH}$ look back 12 months, following Ferson and Khang (2002). The TSA performance generates an HML of 6.05% but the selection component's HML is only 0.77%. Most of the spread in TSA for the CWM is timing. Unfortunately however, only the top quintile shows positive timing ability.

Panel D of Table 4 presents results for the FM measure. Almost all of the FM spread is captured by differences in the AARs. At the fund level, the correlation between FM and its AAR is 0.94. Sorting funds on their FM measures is a lot like sorting them on their AAR. In turn, most of the AAR spread is associated with the passive buy-and-hold drift.

The third quintile in Table 4 contains the median fund sorted on each measure, illustrating the implications of the decomposition on average. The median performance is positive but small under the classical GT and DGTW measures. This is consistent with previoius holdings-based studies, reviewed by Wermers (2000). Decomposing the classical measures, the median AAR is positive for DGTW and FM. The median TSA is negative for every measure except GT. The AAR in excess of buy and hold is negative for the median fund under every measure. This says

[14] In the internet Appendix we examine a sample of index mutual funds, 1992-2012 using DGTW. We find positive and significant AAR in excess of 1% per year, but this is almost entirely driven by the buy-and-hold drift.

that the total before cost performance of the median fund is negative if the passive buy-and-hold drift component, starting with the initial weights, is removed.

In summary, the results from sorting funds on their classical performance measures shows that variation across funds in the differrent measures reflect different abilities. Differences in the classical GT measure reflect mainly TSA. The CWM is 100% TSA, and dominated by timing as opposed to selectivity. Differences in the DGTW and the FM measures reflect mainly AAR, driven mainly by the buy-and-hold drift. The measures agree that the median active performance, net of the passive buy-and-hold drift, is negative.

The Appendix places our decomposition into the broader context of performance attribution and the Internet Appendix presents and empirical example. This shows that for the fund groups of Table 4, the abnormal returns are a small part of total fund returns and confirms that AAR is the larger component for the levels, compared with TSA. A “normal” return plus the AAR captures about 96% of the average return levels and 94% of the cross-sectional variation in fund average returns. The AAR is dominated by the buy-and-hold drift. The TSA is the sum of positive timing and negative selection abilities for each group, suggesting that timing ability might be better than previous studies find. TSA is the main driver of the GT measure, both in the levels and, as shown in Table 4, in the cross-section. The timing component is larger than the average selectivity component of TSA in the GT measure. By focussing on the TSA in the classical measures, timing ability seems more evident. This may be useful in future research.

6.2 Multiple Comparisons

Although Table 4 says that the median fund does not have significant before-cost performance, that does not mean that no funds have ability. Perhaps the extreme performers have true ability.

We address this question by examining the cross-sectional distribution of the performance measures, using simulation to account for the multiple hypothesis tests. To illustrate the issue, imagine that no fund has ability and that a test with a size of 5% is used. We then expect to find that 5% of our sample of 3596 funds, or about 180 funds, would present performance with t-ratios in excess of two.

We conduct simulations under the null hypothesis of no ability. To impose the null hypothesis that TSA is zero, we randomly scramble the returns for a given fund's stocks through time, using the original porfolio weights to measure the performance. To remove an average cross-sectional relation between weights and stocks' alphas, we randomly scramble the returns for a given fund-month across the stocks held that month, and use the original portfolio weights to measure the performance.[15] Each experiment generates a critical value for the performance measure, such that 5% of the estimates under the null exceed the critical value. We are interested to see how often the measures in the original data exceed the critical values. We also use this information to estimate the fractions of funds in the population that have zero, positive or negative true performance.

To evaluate the statistical significance of the extreme performers we model an indicator variable for the event that a measure exceeds a bootstrapped critical value as a binomial random variable, taking the value one with 5% probability and the value zero with 95% probability. The test statistic is the fraction of actual funds whose performance exceeds the critical value. A t-ratio for the significance of this fraction is the difference between the fraction and 0.05, with a standard error of about 4.4%. Thus, significant performance is found when more than about 13.8% of the funds' performance measures exceed the critical value.[16]

[15] The null hypothesis is not that the AAR is zero, but that it is equal to the alpha of the average stock. The AAR is the average stock alpha plus N times the cross-sectional covariance between the average weights and the alphas. We test whether the covariance is zero.

[16] The fraction of the cases exceeding the critical value is $f = (1/n) \sum_i I(x_i=1)$, where $I(x_i=1)$ is the indicator variable

We estimate the fractions of funds in the population with nonzero performance along the lines of Barras, Scaillet and Wermers (BSW, 2010). The approach starts with the empirical p-values for the t-ratios of the performance measures. For each fund, the empirical p-value is the fraction of funds with larger absolute t-ratios (for two-tailed tests) in the simulations under the null hypothesis. Let π_0 be the fraction of zero-performance funds in the population. BSW estimate π_0 following Storey (2002), by choosing a large cutoff p-value, λ, found by minimizing the mean squared error around the smallest π_0 value found for the various cutoff p-values. They find that the estimates are not very sensitive to the choice of λ, provided it is large enough, and get similar results to the optimized value if λ=0.5 or 0.6 is chosen. We try both 0.5 and 0.6 and find little difference in the results. We report result for λ=0.5. Using the cutoff p-value there is a fraction of funds, $F_g(\lambda)$, for which the null is rejected in favor of good performance. There is also a fraction of funds, $F_b(\lambda)$, for which the null is rejected in favor of bad performance. The estimate of π_0 is $[1- (F_g(\lambda) + F_b(\lambda))]/(1-\lambda)$. Using this estimate, the fraction of good performing funds, π_g, is then estimated as: $\pi_g = F_g(\gamma) - \pi_0 (\gamma/2)$, where γ is a different cutoff significance level, also estimated by BSW with a MSE minimization by simulation. They find that similar results are obtained when γ = 0.35 or 0.45 is simply selected. We try different values and the results are little changed. We report results for γ = 0.35.

Table 5 summarizes the results of the multiple comparison analyses. The critical values are reasonably symmetric for the TSA measures, while the net AAR is left-tail skewed under the null hypothesis. The fractions of actual funds exceeding the critical values present evidence that some funds have nonzero performance. According to the GT measure there are some funds with

for the *i*-th measure exceeding the 5% critical value. Under the null the mean of *f* is 5% and its variance, accounting for the correlation across the tests, is .05(.95)[(1/n) + {(n-1)/n}ρ], where ρ is the average correlation of the tests across n funds. We use the estimates of Barras, Scaillet and Wermers (2010) and Chen and Ferson (2020), setting ρ = 0.04.

significant net AAR in each tail, some with significant negative TSA and some with significant negative TSA + net AAR, but there is no evidence of significant positive performance. The DGTW and FM measures present no evidence of significant performance in the tails.

The estimated fractions of funds with nonzero true performance are labelled “Fractions Nonzero.” Less than 11% of the funds are found to have positive performance by any measure, while 26-40% have negative true performance, depending on the measure. More than 50% of the funds are estimated to have zero true performance. The BSW estimate of the fraction π_0 is known to be upwardly biased when the tests have low power (e.g. Storey (2002), Chen and Ferson (2020), Barras Scailett and Wermers, 2019). However, in our setting the fact that we re-use the original weights each period in the simulations biases the fractions of nonzero alphas upwards (Fama and French, 2010), and the two biases are offsetting. Future studies can use more refined techniques, beyond the scope of this paper, to improve on our estimates of these fractions.

6.3 Fund flows

Given the heterogeneity in the measures it is interesting to study the response of investor flows to the various classical measures and their components. The flow is measured following Sirri and Tufano (1998) as the change in total net assets in excess of that implied by a fund’s return. The first set of tests follow Berk and van Binsbergen (BvB, 2016), regressing the signs of annual new money flows on the signs of the performance measures over the past year. The performance measures are estimated and averaged over the past year, using the “up-to-t” version of the Haj10 estimator that uses data no closer than 13 months prior to the flow date.

Table 6 reports slope coefficients and t-ratios for the panel flow-on-performance regressions, with or without a set of fund level control variables. We control for several fund characteristics, including the fund expense ratios and turnover. These latter two are potentially important if investor flows respond to after-cost performance, while holdings-based measures reflect before-cost performance. Similar to previous studies we find that for a given before cost performance, the flow response is less if the fund has a higher expense ratio or a higher turnover, both indicating lower after-cost returns to investors. However, these control variables have a small impact on the coefficients and are not shown in the table.[17] We also report a transformation of the coefficients from BvB, which measures the probability that the signs of the abnormal returns and the subsequent flows match. Details are in the table header. To compare with the earlier studies we include the CAPM in these tests, which we implement as a version of the FM measure with only a single market factor.

In the first panel of Table 6 the original holdings-based measures are used. All of the coefficients have t-ratios larger than 4.5. The probabilities of matching signs are all in the 54-58% range. This is similar to BvB, who found a 57-63% range across a different set of models. Barber, Huang and Odean (BHO, 2016) also conclude that none of the models explain a large fraction of fund flows.

Del Guercio and Tkac (2002), BvB and BHO find that investor flows respond more strongly to relatively simple measures of fund performance. The CAPM is the simplest measure here, but it does not perform the best. All the models are close, but the CAPM beats GT, and the DGTW measure performs the best at predicting flows. The FM measure betas the CAPM but not DGTW. The FM, GT and DGTW measures are not directly examined by BvB or BHO.

Table 6 shows that the TSA components of the measures produce smaller flow responses than

[17] The control variables in Table 6 that present t-ratios above two and their coefficient signs are: the fund expense ratio (<0), turnover (<0), a dummy variable for an aggressive growth style (>0), fund age (<0) and return volatility (<0). The signs all make sense, are consistent with previous research and are similar across the performance measures.

the original measures, with no t-ratio larger than 2.5 and all the fractions correct less than 53%. The AAR components elicit a greater response than the TSA component for GT and DGTW, but a smaller response than TSA for FM and CAPM, where it is not statistically significant.

The striking results appear when we remove the passive buy-and-hold drift from the AAR. The net AAR dominates the TSA for every model, and the t-ratios on net AAR are larger than 4.7 in every case. Even in the FM measure the net AAR performs about as well as GT or DGTW. The net AAR component of performance elicits a strong flow response. The bottom rows of Table 6 show results for the buy and hold drift component of the measures. None of the coefficients are statistically signficant and they are all economically small. It seems that investor flows do not respond to the passive buy-and-hold drift.

The panel regressions in Table 6 capture both time-series and cross-fund variation in investor flows and fund performance. We estimate versions of this table including fun dummy variables, so that the coefficients focus on the time-series covariation between the flows and performance, averaged across funds. Most of the results are little changed, except that the four significant cases using the TSA performance fall out. We also estimate a version of the table with time dummy variables, so that the coefficients focus on the cross-sectional covariation between flows and performance, averaged through time. In this case the original measures appear to be even stronger flow predictors, especially the FM and CAPM which now sport t-ratios larger than ten and fractions correct larger than 60%. In no case does a TSA component of performance now deliver a t-ratio in excess of two. The AAR component is much stronger than before with t-ratios now larger than 4.5 even for the FM and CAPM. The results for the net AAR are quite similar to those in Table 6, although the buy and hold drift component becomes significant for the FM measure where it was not before. Overall, these results say that investor flows respond to the classical measures, especially their differences across funds, and that very little of this has to do with the TSA component of performance. The AAR and especially the AAR net of the passive buy-and-hold drift, are strong predictors of investor flows into mutul funds.

We also conduct a flow-performance analysis similar to Sirri and Tufano (1998). Each performance measure is used to form ranked performance within each of five performance quintiles. A piecewise linear regression allows for a different slope, measuring a different sensitivity of flows to the performance, in each performance range. We estimate models with and without controls for other fund characteristics and report the tables in the Internet Appendix.

Ippolito (1992) finds a negative flow response to below average performance and a positive response to above average performance. Sirri and Tufano (1998) find a weak negative or zero coefficient for the low performance quintile, mixed signs in the intermediate quintiles and strong positive slopes in the high performance quintiles under various measures of performance. Our findings using the classical holdings-based measures are roughly similar. Five of the 15 slopes for the three models (GT, DGTW and FM) and five quintiles have absolute t-ratios larger than two. Using the AARs as the performance measures nine of the 15 slope coefficients have t-ratios larger than two. The t-ratios for flow on performance in the top quintile of AARs is 4.86 for the DGTW measure.[18] Using the AARs net of buy-and-hold drift, the results are even stronger and twelve of the 15 slopes have t-ratios above two. We conclude that the AARs net of the bias from buy-and-hold drift elicit the strongest overall flow response.

6.4 Control Variables

This section illustrates the use of control variables in the panel regressions. The regression with a

[18] We also examine regressions where the AAR is averaged over the previous three years, which produces similar but less statistically significant results.

control variable, X^i_t is:

$$r^i_{t+1} = a^i + \beta\, w^i_t + \gamma\, X^i_t + \varepsilon^i_{t+1}. \qquad (18)$$

The slope coefficient, β, is the conditional TSA given X, measuring the information in the weights about the future unexpected returns given the stock characteristics, X^i_t. The classical CWM is a conditional performance evaluation measure, but it conditions on macroeconomic information. We use the dividend yield of the stock as the control variable to illustrate.

The use of stock level control variables in the panel provides a simple way to ask if fund performance is "explained" by the characteristics of the stocks held by funds. This is a classic question in the literature. For example, Carhart (1997) and Wermers (2003) argue that much of the persistence in fund performance is explained by their holdings of momentum stocks. Zheng (1999) and Sapp and Tiwari (2004) argue that the "smart money" effect of Gruber (1996), where new money flows seem to earn higher returns in funds, is mainly due to the momentum effect in the stocks held. Sheng, Simutin and Zhang (2019) explain the relation between expense ratios and fund performance using the two new factors in the Fama and French (2015) five-factor model. Matalin-Saez (2020) argues that the relation of fund performance to their factor model R-squares (Amihud and Goyenko, 2013), is explained by the holdings of low R-square stocks. Table 7 joins this literature by showing that the fund level dividend yield relation to performance is significantly reduced by controlling for the dividend yields of the stocks held.

The Frisch-Waugh (1933) theorem applies to the estimation of Equation (18). We first demean the variables, accounting for the stock fixed effects. In the second step we regress the demeaned returns on the demeaned X^i_t and take the residuals, vr^i_{t+1}. We also regress the demeaned weights on the demeaned X^i_t and take the residuals, vw^i_t. Let $u_r = Vec(r^i_{t+1}\text{-}E(r^i))$, $u_w =$

$\mathrm{Vec}(w^i_t\text{-}E(w^i))$, $u_x = \mathrm{Vec}(X^i_t\text{-}E(X^i))$, where the expectations are estimated using the sample means. The second step regressions are $u_r = \beta_r\, u_x + v_r$ and $u_w = \beta_w\, u_x + v_w$. The third step is to regress v_r on v_w to estimate the coefficient, β, in Equation (18). With scaling we regress $(v_w' v_w/T)\, v_r$ on v_w. The OLS coefficient is $(v_w' v_w/T)^{-1}\, (v_w' v_r/T)\, (v_w' v_w/T)$, which converges in probability to $E((u_r - \beta_r\, u_x)'(u_w - \beta_w\, u_x)) = E(u_w'\, (u_r - \beta_r\, u_x)) = \mathrm{Cov}(\, w';\, (r–E(r|X))$.

With the Haj10 and DiffIV methods, we use instrumental variables instead of OLS, where the rolling demeaned variables are the instruments. Giles (1984) shows that the Frisch Waugh (1933) theorem holds for instrumental variables estimation.

Most of the earlier studies use a different approach. Factors are commonly formed from portfolios of characteristic-sorted stocks, the high minus low portfolio return differences are the new factors, and returns are adjusted using the new factors. For example, Carhart (1997) produces the famous four-factor model, where the new momentum factor is the return of high past return stocks less the return of low past return stocks.

A high-low portfolio is similar to a cross-sectional regression coefficient of returns on the characteristic (see Fama (1976), or Ferson, Sarkissian and Simin, 1999). Thus, the factor-adjusted return is similar to a cross-sectional regression version of the second step above. Specifically, with OLS Fama-MacBeth (1973) style cross-sectional regressions, the factor adjusted return for characteristic X^i_t is $r_{it+1} – (\Sigma_i\, r_{it+1}\, u_{xit})/(\Sigma_i\, u_{xit}^2)\, X^i_t$, and $\gamma_{t+1} = \Sigma_i\, r_{it+1}\, (u_{xit}/\Sigma_i\, u_{xit}^2)$ is the factor premium at time *t+1*. This is a portfolio of stock returns with weights equal to $(u_{xit}/\Sigma_i\, u_{xit}^2)$. In the panel the adjusted return is $r_{it+1} – (\Sigma_t\Sigma_i\, r_{it+1}\, u_{xit})/(\Sigma_t\Sigma_i\, u_{xit}^2)\, X^i_t$. The panel coefficient in (18) is $\gamma = \Sigma_t\, \gamma_{t+1}\, [\, (\Sigma_i\, u_{xit}^2)/(\Sigma_t\Sigma_i\, u_{xit}^2)]$, a weighted average of the Fama-MacBeth risk premiums over time. The weights are higher for months with greater cross-sectional variance in the X^i_t. Ferson and Harvey (1999) propose a similar weighting for improved efficiency in the estimation of the average premium.

We use the dividend yield of a stock as the X variable in Table 7. The Brennan (1970) after-tax CAPM predicts a positive premium associated with a stock's dividend yield. The premium may not be captured by the factor models and therefore may appear in the stocks' returns and risk-adjusted returns.[19] If there is a premium for dividend yield in the stock returns the models leaving it out could confuse the dividend premium with investment ability. We find correlations across the stocks in our sample, between their alphas and their dividend yields, of 0.074 in the DGTW model and 0.258 in the Carhart (1997) model. Controlling for the dividend yield, we measure performance that is not associated with a market premium for holding high yield stocks.

Table 7 shows that for the median dividend yield fund, controlling for stocks' dividend yields has little influence on the slope coefficients, except in the DGTW measure where there is a large statistical impact. There is little impact on the DGTW TSA, so controlling for stocks' dividend yields mainly reduces the AAR component of the DGTW measure. The right hand columns of the table, labelled HML, ask if there are differential effects of the control variable for the high versus low dividend yield funds. The dividend yield control reduces the GT measure more for the high-yield funds. The dividend yield control variable also removes the superior performance of the high-yield funds under the original FM and DGTW measures.

5.6 Persistence

Early studies of fund performance persistence (e.g., Carlson (1970), Hendricks, Patel and Zeckhauser, 1993) found evidence of "hot hands," where past relatively good fund performance predicts good future performance. The early literature confronts the effects of data base biases

[19] The empirical evidence on the cross-sectional relation between stock returns and dividend yields is mixed. See Chen, Grundy and Stambaugh (1990) for evidence and a review of earlier studies.

such as survival bias (Ross, Ibbottson, Goetzmann and Brown (1992), Hendricks, Patel and Zeckhauser (1997), Carpenter and Lynch, 1999). Carhart (1997) revisits performance persistence using an early version of the then newly-developed CRSP mutual fund data base; its premier feature being the attempt to include the dead funds, and thus control survivorship bias. Carhart argued that persistence is related to individual stock momentum in fund holdings, and found that controlling for a momentum factor removes most of the persistence in mutual funds' reported return performance. The remaining persistence concentrates in the poorly performing funds (see also Brown and Goetzmann, 1995). The persistence of poor performance is also found in pension funds by Christopherson, Ferson and Glassman (1998) and in hedge funds by Agarwal and Naik (2000) and others.

The literature distinguishes short term from long term performance persistence. Most of the early studies describe persistence over one to five-year future horizons. Berk and van Binsbergen (2015) find persistence using decile-sorted gross-of-expense ratio alphas, multiplied by the assets under management (TNA), which they denote as value added. Bollen and Busse (2004) use daily fund return data to estimate timing and selectivity measures, and find short-term performance persistence in both measures. Sorting performance into deciles, the top performing decile generates just under 40 basis points of abnormal return over the next quarter, but the performance dissipates quickly in the future.

We examine persistence in performance in Table 8, where funds are sorted into deciles using the up-to-t versions of the holdings-based performance measures, their TSA and net-AAR components. We are curious as to which component is the better predictor. We examine the future performance of the portfolios measured by the DGTW gross alphas in Panel A. Carpenter and Lynch (1999) examine various statistics for measuring persistence using simulations and find that the t-ratio for the extreme decile return differences, HML, is attractive on the basis of power and robustness. The time series averages of the decile HML portfolio excess returns are reported. The t-ratio for the mean is computed following Fama and MacBeth (1973), adjusted for autocorrelation as described previously.

Table 8 finds some persistence in performance when sorting on the original measures. The HMLs after sorting on the original measures are more than 1.6% per year for the first quarter, a similar magnitude as in Bollen and Busse (2004), dissipating over longer horizons, with few significant beyond the first year.

Table 8 shows that sorting on AAR produces similar results as sorting on the original measures. Sorting on the TSA component provides little evidence of persistence, suggesting that the information in TSA is short-lived, relative to the information in AAR.

We also examine future performance as the realized value added measure of Berk and van Binsbergen (2015). These are the DGTW alphas multiplied by the TNA each month, denoted S_{it}. The time series averages of the decile portfolio values, $(1/T)\ \Sigma_i\ S_{it}$ are examined. The t-ratio for the mean is computed following Fama and MacBeth (1973), adjusted for autocorrelation. The results are shown in Panel B of Table 8 in millions of dollars. (We also examine constant dollars using a CPI deflator and the results are similar.) For the first quarter after sorting the HML difference in value added is 0.59 to 0.82 million dollars, but marginally significant only for the GT measure, and dissipating beyond the first quarter. Again, sorting on the AARs is about as informative as sorting on the original measures, and the TSA produces no information about the future performance.[20]

[20] A caveat, most relevant here for the DGTW measure, is that sorting on and measuring the future performance with the same measure can induce spurious persistence if there is a persistent bias in the measure (e.g., a missing dividend yield factor).

6.6 Robustness

This section briefly summarizes a number of robustness checks and additional analyses. More details are in the Internet Appendix. We ask if grouping funds by proxies for active fund management reveals differrences in their TSA, AAR or in their net AAR. Previous studies suggest several proxies for active management. We sort funds each month by the different active management measures, calculate performance for each fund for the next month like in Table 4 and average across the funds in each group. The sorts broadly suggest that active management is much more strongly associated with TSA than with AAR.

We examine how the performance measures are related to other fund characteristics, regressing the monthly measures on a set of lagged fund characteristics. The most striking finding is a strong negative relation of TSA to a measure of a fund's tendency towards a buy-and-hold strategy. We also find a positive relation of performance to fund dividend yield. This appears for all of the measures, but is weaker in the TSA and in GT, indicating that the effect concentrates in the AAR component of performance. This motivates our use of dividend yield as the stock level control variable in Table 7.

We compare the ability of the various holdings-based performance measures and their embedded AARs to detect funds whose holdings pick stocks that will subsequently outperform. For these predictive exercises we estimate the measures for each month using an "up-to-t" version of the Haj10 estimator, so that the performance measure uses only data available before the forecast date. We find no evidence that any of the measures can predict the future stocks' abnormal performance.

Many of the results use monthly stock returns and monthly weights constructed from quarterly holdings data. We argued that the performance measures should be very similar using quarterly or monthly weights, when the monthly weights update the quarterly using returns data. Tables in the internet appendix confirm that this is the case.

Using quarterly data the same price appears in the weight and in the denominator of the future return calculation, so an error in the price might induce a spurious negative relation between the lagged weight and the future return. In momentum studies (e.g. Jegadeesh and Titman, 1993) it is common to skip a day between the formation period and the future return calculation interval to handle such concerns. In a similar spirit the Internet Appendix uses 29-day returns on the left hand side of the regressions, skipping a day relative to the price data in the weights. The DGTW measures are somewhat smaller using the 29-day returns, but otherwise the results are similar.

We also examine the impact of how the returns are demeaned in the Haj10 approach; in particular whether it matters if the fund is actually holding the stock during the period used to demean. In one case demeaning uses only stock-month returns where the fund holds the stock and in the other, the demeaning uses only stock-month returns where the fund does not hold the stock. The results are similar.

7. Conclusions

Holdings-based portfolio performance measures are panel regressions of future stock returns on lagged portfolio weights. We illustrate how clustered standard errors, fixed effects and control variables apply. The panel regression faces a lagged stochastic regressor bias, similar to that described by Stambaugh (1999) and Hjalmarsson (2008, 2010), when stock fixed effects are included. We evaluate several ways to address the bias using simulations, and find two bias correction methods that work well.

Stock fixed effects in the model isolate the time-series predictive ability in portfolio weights (TSA) from the average ability (AAR). TSA is the ability of current portfolio weights to predict future returns. AAR appears when the long term asset allocation policy results in portfolio weights whose average values are higher for higher alpha stocks. Funds that use dynamic strategies, such as market timing funds, likely focus more on TSA performance. Quantitative portfolios with a policy of sorting on stocks' characteristics at low cost may have greater AAR. Differences in the classical performance measures across stocks are mainly driven by the TSA for some measures and by the AAR for others, suggesting that in practice users may wish to focus on different measures for different types of funds.

We illustrate the use of control variables in the performance regressions. Stock level control variables provide a natural way to address a common question in the literature, as many studies ask to what extent fund performance is explained by the characteristics of the funds' holdings (e.g., Carhart (1997), Sapp and Tiwari (2004) and others cited above). Our example joins this literature by showing that the fund level relation of performance to dividend yield is at least partially controlled by the dividend yields of the stocks held.

AAR contains an effect we call the buy-and-hold drift. Buy-and-hold weights drift towards the stocks with higher average returns during an evaluation period. Because average returns and alphas in the popular models are highly correlated across stocks, buy and hold weights drift towards the high alpha stocks, and this shows up as positive AAR. AAR appears in simulated buy-and-hold and momentum strategies, and also appears in index funds. We remove the passive buy-and-hold drift from the AAR component of the classical performance measures, and we find that the active net AAR elicits a stronger new money flow response from fund investors and better predicts future fund returns.

In addition to our methodological contributions, we contribute to the empirical evidence on performance using holdings-based performance measures. We find that the before cost TSA is not positive for the average equity fund. The net AAR, after adjustment for the passive buy-and-hold drift, is also zero or negative.

Future research can use our methods to refine the evidence based on holdings-based performance measures. For example, it should be interesting to study TSA and AAR in the performance of hedge funds, bond funds, international mutual funds, and other managed portfolios for which holdings data are available. The distinction between TSA and AAR provides a window into short term versus long term fund investment decisions. Future studies can use our insights to explore this and other issues.

Appendix

A.1 Components of Performance

Brinson, Hood and Bebower (BHB, 1986) provide a decomposition of pension fund returns using the asset classes stocks, bonds and cash. Each asset class *i* has a policy return, r_{iB}. For the equity style mutual funds examined here, it makes sense to interpret asset classes in terms of investment style, and we use the equity style boxes of DGTW. There are 125 value-weighted, and thus passive returns in the (5x5x5) styles. This can be represented as an N-vector, where each of the N stocks held by a fund is assigned one of the style benchmark returns, $r_{Bt+1} = \mathbf{r}_{t+1}{}^{D}$. BHB describe the actual portfolio weights of a fund in style *i*, which we denote $w_{ait} = \Sigma_{j\varepsilon i}\ w_{jt}$, where w_{jt} is the weight in stock *j* at time *t*, and the summation is over the stocks in style *i*. BHB describe a fund's policy weight, w_{iB}, for asset class *i*, as the constant full-sample (10-year) average weight of the fund's actual weights in the class, $w_{iB} = (1/T)\Sigma_t\ w_{ait}$, or in population $w_{iB} = E(w_{ai})$. The policy return of the fund is then $r_{Bt+1} = \Sigma_i\ w_{iB}\ r_{iBt+1}$. The fund earns an actual (active) return in each asset class, $r_{ait+1} = \Sigma_{j\varepsilon i}\ w_{jt}\ r_{jt+1} / \Sigma_{j\varepsilon i}\ w_{jt}$, where we normalize the weights in the stocks in each class to sum to one. The other components of a fund's performance in BHB are:

$$\text{Timing: } \Sigma_i\ (w_{ait} - w_{iB})\ r_{iBt+1} = \Sigma_i\ (w_{ait} - E(w_{ai}))\ r_{iBt+1} \qquad \text{(A.1)}$$

$$\text{Selection: } \Sigma_i\ w_{iB}\ (r_{ait+1} - r_{iBt+1})$$

The expected value of the BHB timing term is TSA applied at the asset class level, a version of benchmark or style timing. The expected value of the selection term corresponds to $E(\mathbf{w})'\boldsymbol{\alpha}$, or our AAR, at the asset class level.

BHB find that the normal or policy return, $\Sigma_i\ E(w_{ai})\ r_{iBt+1}$, applying the average weights to the passive policy returns, represents the lion's share of fund portfolios' average returns and variances of quarterly returns. However, Jahnke (1997) points out that the cross-sectional variance of the policy returns is much smaller than the cross-sectional variance of funds' realized returns, so the policy return does not explain much of the cross-section of realized returns. Indeed from Brinson, Singer and Beebower (BSB 1991, Table III) the ratio of the cross-sectional variances of policy returns to total returns is only $(0.49\%/1.75\%)^2 = 7.8\%$. From BSB (Table 6) the ratio of the cross-sectional variances of policy returns to total returns is only $(0.22\%/1.43\%)^2 = 2.4\%$. However, the sum of the policy return and the selection return does explain the lion's share of the cross-sectional variance in fund returns. The timing component is relatively small. Again from BSB (Table III) the ratio of the cross-sectional variances of policy returns and selection returns to total returns is $(1.66\%/1.75\%)^2 = 89.9\%$. From BSB (Table 6) the ratio of the cross-sectional variances of policy returns plus selectivity returns to total returns is $(1.33\%/1.43\%)^2 = 86.5\%$.

In summary, the evidence of BHB, BSB and Janke (1997) suggests that the average magnitude of E(w'r) is close to $E(w)'E(r_B)$, the normal or policy return. Janke's interpretation says that the cross sectional variation in E(w'r) is mostly captured by the cross-sectional variation in selection plus policy. The expected value of BHBs selection return corresponds to AAR at the asset class level.

In DGTW (1997) the holdings-based portfolio return, $\boldsymbol{w}_t'\mathbf{r}_{t+1}$, is decomposed into the sum of three terms, denoted as Characteristic Selectivity (CS), Average Style (AS) and Characteristic Timing (CT). We study the most popular CS term in our paper, denoting it as the DGTW measure. DGTW (1997) find little evidence that funds have nonzero CT. This section describes how our TSA and CSA are related to the three DGTW measures. Their definitions, using our notation, are:

$$\mathrm{CS}_t = \boldsymbol{w}_t'(\mathbf{r}_{t+1} - \mathbf{r}_{t+1}^{D(t)}), \tag{A.2}$$

$$\mathrm{CT}_t = \boldsymbol{w}_t'\mathbf{r}_{t+1}^{D(t)} - \boldsymbol{w}_{t-k}'\mathbf{r}_{t+1}^{D(t-k)},$$

$$\mathrm{AS}_t = \boldsymbol{w}_{t-k}'\mathbf{r}_{t+1}^{D(t-k)}$$

where $\mathbf{r}_{t+1}^{D(j)}$ is the vector of DGTW benchmark returns, when the benchmarks for the stocks are assigned at time *j*. The sum of the three components is $\boldsymbol{w}_t'\mathbf{r}_{t+1}$, and the time-series averages of the components are reported.

A.2. Relation Among the Components

The BHB and DGTW decompositions are similar but not identical. If we interpret the lagged weights and $\mathbf{r}_{t+1}^{D(t-k)}$ in DGTW as proxies for the policy weight and return, then the BHB timing measure corresponds to the DGTW characteristic timing. Under that interpretation the DGTW AS measure corresponds to the policy return in BHB.

DGTW assume that the portfolio weights at time *t-k* are independent of returns at time *t+1*, where *k*=12 months. If we also assume that the assignment of a stock to one of the 125 benchmark portfolios is relatively stable from year to year, so that D(t) ≈ D(t-k), then from (A.2) we can write $E(CT) = E(\boldsymbol{w}_t'\mathbf{r}_{t+1}^{D(t)}) - E(\boldsymbol{w}_{t-k}'\mathbf{r}_{t+1}^{D(t-k)}) = Cov(\boldsymbol{w}_t'; \mathbf{r}_{t+1}^{D(t)}) - Cov(\boldsymbol{w}_{t-k}'; \mathbf{r}_{t+1}^{D(t-k)})$, as the two $E(\boldsymbol{w})'E(\mathbf{r}^D)$ terms cancel. If the weights at *t-k* are unrelated to returns at *t+1* the second covariance is zero, and we have $E(CT) = Cov(\boldsymbol{w}_t'; \mathbf{r}_{t+1}^{D(t)})$. This is a version of benchmark timing, and corresponds to the BHB timing return.

The selectivity components in the studies differ. The DGTW CS measure uses the actual fund weight while the BHB selectivity measure uses the expected weight. The expected value of the BHB selectivity measure is AAR at the asset class level. The difference between the expected values of these two measure is a covariance between the actual weights and the future actual returns in excess of policy returns. This covariance is the TSA for the abnormal returns. So, the DGTW CS measure includes TSA + AAR, while the BHB selectivity measure is all AAR.

References

Agarwal, and Naik, 2000, Multiperiod performance persistence analysis of hedge funds, Journal of Financial and Quantitative Analysis 35, 327-

Amihud, Yakov and Rusian Goyenko, 2013, Mutual Fund R^2 as Predictor of Performance, *Review of Financial Studies* 26, 667-695.

Amihud, Yakov and Clifford Hurvich, 2004, Predictive Regressions: A Reduced-Bias Estimation Method, *Journal of Financial and Quantitative Analysis* 39, 813-841.

Amihud, Yakov, Cliffird Hurvich and Yi Wang, 2008, Multiple-Predictor Regressions: Hypothesis Testing, *Review of Financial Studies* 22, 414-434.

Amihud, Yakov, Clifford Hurvich and Yi Wang, 2010, Predictive Regression with Order-p Autoregressive Predictors, *Journal of Empirical Finance* 17, 513-525.

Anderson, T.W and Cheng Hsiao, 1981, Estimation of dynamic models with error components, *Journal of the American Statistical Association*, 589-606.

Barras, L., Scaillet, O., & Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance*, *65*(1), 179-216.

Barras, L., Scaillet, O., & Wermers, R., 2019. Reassessing False discoveries in mutual fund performance: Skill, luck or a lack of power? *Swiss Finance Research Institute*, working paper no. 19-61.

Barber, Bradford, X. Huang and Terry O'Dean, 2016, Which risk factors matter to investors? Evidence from mutual fund flows, *Review of Financial Studies* 29, 2600-2642.

Becker, C., Ferson, W., Myers, D., Schill, M., 1999. Conditional Market timing with Benchmark investors. *Journal of Financial Economic*s 52, 119-48.

Berk, Jonathan, and Richard Green, 2004, Mutual fund flows and performance in rational markets, Journal of Political Economy 112, 1269-1295.

Berk, Jonathan and Jules vanBinsbergen, 2016, Assessing asset pricing models using revealed preference, *Journal of Financial Economics* 119, 1-23.

Berk, Jonathan and Jules van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1-20.

Bollen, Nicolas and Jeffrey Busse, 2004, Short-term persistence in mutual fund performance, *Review of Financial Studies* 18, 569-

Brennan, Michael J., 1970, Taxes, market valuation and corporate financial policy. *National tax* *journal* 23, 417-427.

Brinson, Gary P., R. Hood and Gilbert L. Beebower, 1986, Determinants of portfolio performance, *Financial Analysts Journal* /January-February, 133-138.

Brinson, Gary Brian Singer and Gilbert L. Beebower, 1991, Determinants of portfolio performance II: An update, *Financial Analysts Journal* /May-June, 40-48.

Brown, Stephen J. and William N. Goetzmann, 1995, Performance Persistence, *Journal of Finance* 50, 679-698.

Brown, S.J., Goetzmann, W., Ibbotson, R.G. and Ross, S.A., 1992. Survivorship bias in performance studies. *The Review of Financial Studies*, *5*(4), pp.553-580.

Busse, Jeffrey and Qing Tong, 2012, Mutual fund industry selection and persistence, Review of Asset Pricing Studies 2, 245-272.

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

Carlson, Robert S., 1970, The aggregate performance of mutual funds, *Journal of Financial and Quantitative Analysis* 5, 1-32.

Carpenter, Jennifer N. and Anthony W. Lynch, 1999, Survivorship bias and attrition effects in measures of performance persistence, *Journal of Financial Economics* 54, 337-374.

Chen, N.F., Grundy, B. and Stambaugh, R.F., 1990. Changing risk, changing risk premiums, and dividend yield effects. *Journal of Business*, pp.S51-S70.

Chen, Yong, and Wayne Ferson, 2020, How many good and bad funds are there, Really? Chapter 108 in the *Handbook of Financial Econometrics, Statistics and Technology*, World Scientific Press. C.F.Lee, editor (forthcoming).

Christopherson, Jon, W. Ferson and D. Glassman, 1998, Conditioning manager alphas on economic information: Another look at the persistence of performance." *The Review of Financial Studies* 11, 111-142.

Copeland, T.E. and David Mayers, 1982, The value line enigma (1965-1978): A case study in performance evaluation, *Journal of Financial Economics* 10, 289-321.

Cornell, B., 1979, Asymmetric Information and portfolio performance measurement, *Journal of Financial Economics* 7, 381-391.

Cremers, Martijn and Antii Petajisto, 2009, How active is your fund manager? A new measure that predicts performance, *Review of Financial Studies* 22, 3329-3265.

Cremers, Martijn, Antii Petajisto and Eric Zitwewitz, 2012, Should benchmark indices have alpha? Revisiting Performance evaluation, *Critical Finance Review*.

Dahlquist, M., Engström, S. and Söderlind, P., 2000. Performance and characteristics of Swedish mutual funds. *Journal of Financial and quantitative Analysis*, *35*(3), pp.409-423.

Daniel Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring Mutual Fund Performance with Characteristic Based Benchmarks, *Journal of Finance* 52, 1035-1058.

Del Guercio, Diane and Paula Tkac, 2002, The determinants of the flow of funds of managed portfolios: Mutual funds versus pension funds, *Journal of Financial and Quantitative Analysis* 37, 523-57.

Doshi, Hitesh, Redouane Elkamhi and Mikhail Simutin, 2015, Managerial activeness and mutual fund performance, *Review of Asset Pricing Studies* 2, 156-184.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 2011, Holdings data, security returns, and the selection of superior mutual funds, *Journal of Financial and Quantitative Analysis* 46, 341-367.

Evans, Richard, 2010, Mutual fund incubation, *Journal of Finance* 65, 1581-1611.

Fama, Eugene F., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance* 25, 383-417.

Fama, Eugene F., 1976, *Foundations of Finance*, Basic Books, New York.

Fama, Eugene F., and Kenneth R. French. "A five-factor asset pricing model." *Journal of financial economics* 116, no. 1 (2015): 1-22.

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

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests. *Journal of political economy* 81, 607-636.

Farnsworth, H., Ferson, W. E., & Jackson, D. S. Todd, 2002, Performance evaluation with stochastic discount factors. *Journal of Business 75*, 473-505.

Ferson, W.E. and Harvey, C.R., 1999. Conditioning variables and the cross section of stock returns. *The Journal of Finance*, *54*(4), pp.1325-1360.

Ferson, Wayne and Kenneth Khang, 2002, Conditional Performance Measurement Using Portfolio Weights: Evidence for Pension Funds, *Journal of Financial Economics* 65, 249-282.

Ferson, Wayne and Jerchern Lin, 2014, Alpha and performance measurement: the effects of investor disagreement and heterogeneity, *The Journal of Finance*, *69*, 1565-1596.

Ferson, Wayne and Haitao Mo, 2016, Measuring Performance with Market and Volatility timing and Selectivity, *Journal of Financial Economics* 121, 93-110.

Ferson, W.E. and Schadt, R.W., 1996. Measuring fund strategy and performance in changing economic conditions. *The Journal of finance*, *51*(2), pp.425-461.

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

Ferson, Wayne, Timothy Simin and Sergei Sarkissian, 2003, Spurious regressions in Financial Economics, *Journal of Finance* 58, 1393-1414.

Ferreira, M. A., Aneel Keswani, Antonio Miguel and Sofia Ramos, 2013, The determinants of mutual fund performance: A cross-country study, *Review of Finance* 17, 483-525

Frisch, Ragnar; Waugh, Frederick V., 1933, Partial Time Regressions as Compared with Individual Trends, *Econometrica 4, 387–401.*

Giles, D.E., 1984, Instrumental variables regressions involving seasonal data, *Economics Letters* 14, 339-343.

Grinblatt Mark, and Sheridan Titman, 1989a, Portfolio Performance Evaluation: Old issues and new insights, *Review of Financial Studies 2, 393-421*.

Grinblatt Mark, and Sheridan Titman, 1989b, Mutual Fund Performance: An Analysis of quarterly portfolio holdings, *Journal of Business 63, 393-416*.

Grinblatt Mark, Sheridan Titman, 1993, Performance measurement without benchmarks: an examination of mutual fund returns, *Journal of Business* 60, 97-112.

Grinblatt, Mark, Sheridan Titman and Russ Wermers, 1995, Momentum investment strategies, portfolio performance and herding: A study of mutual fund behavior, *American Economic Review* 85, 1088-1105.

Gruber, M.J., 1996. Another Puzzle: The Growth of Actively Managed Mutual Funds, Presidential address presented at the American Finance Association, San Francisco, January 1996. *Journal of Finance*.

Hansen, Lars P., 1982, Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pp.1029-1054.

Hendricks, Darryll, Jayendu Patel and Richard Zeckhauser, 1997, The J-shape of performance persistence given survivorship bias, *Review of Economics and Statistics* 79, 161-

Hendricks, Darryll, Jayendu Patel and Richard Zeckhauser, 1993, Hot Hands in mutual funds:

Short-run persistence of relative performance, 1974-1988, *Journal of Finance* 48, 93-130.

Hjalmarsson, Erik, 2010, Predicting Global Stock returns, *Journal of Financial and Quantitative Analysis* 45, 49-80.

Hjalmarsson, Erik, 2008, The Stambaugh Bias in Panel Predictive Regressions, *Finance Research Letters* 5, 47-58.

Ippolito, Richard A., 1992, Consumer reaction to measures of poor quality: Evidence from the mutual fund industry. *The Journal of Law and Economics* 35, 45-70.

Jahnke, William, 1997, The asset allocation hoax, *Journal of Financial Planning*, February/109-113.

Jegadeesh, N. and S. Titman, 1993, Returns to buying winners and selling losers: Implicataions for stock market efficiency, *Journal of Finance* 48, 65-91.

Jiang, George, Tong Yao and Tong Yu, 2007, Do Mutual Funds Time the Market? Evidence from Portfolio Holdings, *Journal of Financial Economics* 86, 724-758.

Kacperczyk, M., Sialm, C. and Zheng, L., 2005. On the industry concentration of actively managed equity mutual funds. *The Journal of Finance*, *60*(4), pp.1983-2011.

Kacperczyk, M., Nieuwerburgh, S.V. and Veldkamp, L., 2014. Time-varying fund manager skill. *The Journal of Finance*, *69*(4), pp.1455-1484.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng, 2006, Unobserved actions of mutual funds, *Review of Financial Studies* 21, 2379-2416.

Kothari, S.P. and Jerold Warner, 2001, Evaluating Mutual Fund Performance, *Journal of Finance* 56, 1985-2010.

Magkotsios, Georgios, 2018, Industry level returns to scale and investor flows in asset management, unpublished PhD dissertation, University of Southern California.

Matalin-Saez, Juan Carlos, 2020, Better performance of mutual funds with lower R2's does not suggest that active management pays, working paper, Universitat Jaume I.

Newey, Whitney and K. West, 1987, A simple, autocorrelation and heteroskedasticity consistent covariance matrix, Econometrica

Pastor, Lubos and Robert Stambaugh, 2009, Predictive Systems: Living with Imperfect Predictors, *Journal of Finance* 64, 1583-1628.

Pastor, Lubos, Robert F. Stambaugh and Lucian Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23-45.

Petersen, Mitchell, 2009, Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches, *Review of Financial Studies* 22, 435-480.

Sapp, T. and Tiwari, A., 2004. Does stock return momentum explain the “smart money” effect?. *The Journal of Finance*, *59*(6), pp.2605-2622.

Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), pp.425-442.

Sheng, J., Simutin, M. and Zhang, T., 2019. Cheaper is not better: On the superior performance of high-fee mutual funds. *Rotman School of Management Working Paper*, (2912511).

Skripnik, Roman, 2019, Mutual fund screening versus weighting, working paper, University of Southern California.

Sirri, Eric and Peter Tufano, 1998, Costly search and mutual fund flows, *Journal of Finance* 53, 1589-1622.

Stambaugh, Robert, 1999, Predictive regressions, *Journal of Financial Economics* 54, 315-421.

Storey, J.D., 2002, A direct approach to false discovery rates, *Journal of the Royal Statistical Society* 64, 479-498.

Wermers, Russ, 2003, Is money really smart? New evidence on the relation between mutual fund flows, manager behavior and performance persistence, working paper, University of Maryland.

Wermers, Russ, 2006, Performance Evaluation with Portfolio Holdings Information, *North American Journal of Economics and Finance* 17, 207-230.

Wermers, Russ, 2000, Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs and expenses, *Journal of Finance* 60, 1655-1695.

Wang, Junbo L., 2015, Can weight-based measures distinguish between informed and uninformed fund managers, PhD dissertation, University of Southern California.

Zheng, L., 1999. Is money smart? A study of mutual fund investors' fund selection ability. *the Journal of Finance*, *54*(3), pp.901-933.

Zhu, Min, 2018, Informative fund size, managerial skill and investor rationality, *Journal of Financial Economics* 130, 114-134.

Table 1: Summary Statistics

Panel A: Descriptive statistics at the fund level

	Mean	Std	Max	Min
TNA ($million)	684	2102	44496	1.03
Number of stocks	114	198	3317	10
Return Gap (%)	0.03	0.23	1.85	-2.60
Active Weight	0.40	0.10	0.90	0.01
Return σ	0.01	0.01	0.21	0.00
R-squares	0.87	0.14	0.99	0.06
ρ_1	0.94	0.04	1.00	0.20
ρ_2	0.89	0.07	1.00	-0.03
ρ_3	0.83	0.11	1.00	-0.53
ρ_4	0.78	0.13	1.00	-0.52
ρ_5	0.74	0.15	1.00	-0.52
Error Corr	0.00	0.02	0.24	-0.19

Panel B: AR coefficients of the weights at the stock level

	Mean	Std	Max	Min
ρ_1	0.90	0.04	1.00	0.26
ρ_2	0.81	0.07	0.99	0.20
ρ_3	0.73	0.09	0.99	-0.55
ρ_4	0.66	0.10	0.99	-0.50
ρ_5	0.59	0.12	0.99	-0.39

In panel A, for each fund in the sample we compute the time-series average of its total net assets (TNA) in millions of dollars, the average number of stocks held, the return gap and the active weight. We compute the sample standard deviations of the monthly reported fund returns, σ, and the R-squares of monthly factor model regressions in the Carhart (1997) four-factor model. We also compute the sample autocorrelations of the funds' monthly portfolio weights, ρ_j, averaged across the holdings, at various lags j, j=1,…5. The means, std, max and min are taken across the funds in the sample. Error Corr is the correlation between the errors of the portfolio weight autocorrelation regression and the panel regression of future stock returns on the weights, averaged across the holdings. The sample period is from 1984 to 2012, and the number of funds is 3596. In Panel B, we compute the autoregression for the weights for each stock and average across the funds that hold the stock. The descriptive statistics are calculated at the stock level.

Table 2: Illustration of Buy-and-Hold Drift

Panel A: Stocks held by Mutual Funds

	DGTW	FM	CAPM	C4
Beginning Weights:				
Highest Alphas	3.87	2.98	2.23	3.84
Lowest Alphas	3.99	2.20	3.67	2.57
Ending Weights:				
Highest Alphas	2.29	2.45	3.57	3.23
Lowest Alphas	0.84	1.16	0.96	1.25
Average Weights:				
Highest Alphas	2.54	3.06	3.15	3.67
Lowest Alphas	1.62	0.91	0.95	0.93
AAR	0.28	0.42	0.95	0.93

Panel B: Stocks of the Standard and Poors 500 Index

	DGTW	FM	CAPM	C4
Beginning Weights:				
Highest Alphas	2.71	4.65	3.52	4.2
Lowest Alphas	2.13	2.73	4.57	3.97
Ending Weights:				
Highest Alphas	8.02	4.27	4.64	5.74
Lowest Alphas	1.19	2.62	2.14	2.07
Average Weights:				
Highest Alphas	3.32	5.29	5.03	5.58
Lowest Alphas	2.13	2.95	3.71	3.64
AAR	0.25	0.36	0.90	0.89

This table records the portfolio weights on the 25 highest and lowest-alpha stocks for buy-and-hold strategies. In Panel A the strategy holds all stocks held by a sample of mutual funds, starting with aggregate total net asset value weights. In Panel B it starts with the stocks in the Standard and Poors 500 index, beginning with the market capitalization weights. The beginning date is January of 1980. The ending date is December of 2012. Different models for alpha are the columns, including models from Daniel, Grinblatt, Titman and Wermers (DGTW, 1997), the stochastic discount factor model of Ferson and Mo (FM, 2016), the Capital Asset Pricing Model of Sharpe (CAPM, 1964) and the four-factor model of Carhart (C4, 1997). The units are annual percent. AAR is E(**w**)'$\boldsymbol{\alpha}$ for the buy-and-hold strategies.

Table 3: Simulations to Address Lagged Stochastic Regressor Bias

----

Panel A: Alternative Holdings-based Performance Estimators in Predictive Panels

True	OLS	OLS with intercept	Haj2008	Within	Diff	Diff IV	Haj10
-0.29	0.13	-0.16	-0.25	-0.37	-0.44	-0.30	-0.30
-0.19	0.17	-0.01	-0.14	-0.32	-0.22	-0.22	-0.20
0.06	0.08	0.04	0.17	-0.02	0.00	0.04	0.05
0.40	0.48	0.39	0.42	0.37	0.35	0.40	0.41
0.58	0.92	0.58	0.60	0.46	0.40	0.60	0.58

Panel B: Accuracy in Estimating AAR

DGTW	DGTW	FM	FM	GT	GT
True AAR	Simulated	True AAR	Simulated	True AAR	Simulated
0.54	0.57	-0.31	-0.28	0.03	0.02
0.39	0.35	-0.14	-0.13	0.03	0.02
0.22	0.26	0.39	0.43	-0.02	0.01
0.21	0.17	0.46	0.49	-0.05	-0.02
0.47	0.40	0.76	0.84	-0.05	-0.04

The table reports the average across 1000 simulation trials, of estimated holdings based performance measures. The units are percent excess return per quarter. True denotes the actual values of the measures used to calibrate the simulations, bootstrapping from versions of Equations (9) and (12). The OLS estimates are the slope coefficients in the baseline panel regression of (7), Diff IV is the differenced IV estimator, Diff is the classical difference estimator of the regression (9) with stock dummies and Within is the classical within-group (least squares with dummy variables) estimator. OLS with intercept is based on a model with a common intercept. Haj10 is the bias adjusted estimator of Hjalmarsson (2010) and Haj2008 is the bias adjusted estimator of Hjalmarsson (2008). GT is the weight change measure of Grinblatt and Titman (1993), DGTW is the measure of Grinblatt, Titman and Wermers (1997). CWM is the conditional weight measure of Ferson and Khang (2002). FM is the stochastic discount factor measure of Ferson and Mo (2016). In panel B the "true" parameter values are the estimates of AAR from single-fund simulations assuming all the stocks exist in each period. These are the differences between the OLS coefficients and the Haj2010 coefficients. The average of the AAR estimates from 1,000 bootstrap simulations, resampling from these data, are shown in each "simulated" column.

Table 4: Cross-sections of Holdings-based Performance Measures

Panel A: Grinblatt Titman Portfolio Change Measure

Fund Quintile:	GT	TSA	AAR
Low	-3.49	-2.34	-0.98
2	-0.72	-0.71	-0.14
Median	0.19	0.10	-0.07
4	1.20	1.04	-0.08
High	4.19	4.15	-0.23
HML	7.69	6.48	0.74

Panel B: DGTW Characteristic Selectivity Measure

Fund Quintile:	DGTW	TSA	AAR	BH	Net AAR
Low	-3.65	-1.38	-2.27	-1.14	-1.13
2	-0.80	-1.03	0.23	0.62	-0.38
Median	0.04	-0.67	0.70	0.85	-0.14
4	0.82	-0.64	1.46	1.29	0.17
High	2.97	0.02	2.95	2.42	0.52
HML	6.61	1.40	5.21	3.56	1.65

Panel C: Conditional Weight-based Measure

Fund Quintile:	TSA	DSel	Timing
Low	-2.93	-0.07	-1.28
2	-0.87	0.28	-0.56
Median	-0.07	0.19	-0.11
4	0.63	0.50	-0.01
High	3.13	0.70	0.79
HML	6.05	0.77	2.07

Table 4, page 2

Panel D: Ferson and Mo Stochastic Discount Factor Measure

Fund Quintile:	FM	TSA	AAR	BH	Net AAR
Low	-19.36	-2.19	-17.18	-15.87	-1.31
2	-5.29	-2.46	-2.83	-2.10	-0.73
Median	-1.30	-1.70	0.41	1.47	-1.07
4	1.56	-1.33	2.89	3.76	-0.87
High	8.36	-1.26	9.62	10.10	-0.47
HML	27.72	0.92	26.80	25.97	0.83

The various performance measures are estimated for each fund using a panel regression and its full sample. GT is the portfolio change measure, DGTW is the DGTW characteristic selectivity measure, FM is the Ferson and Mo SDF-based measure and CWM is the conditional weight based measure. Funds are sorted into five groups on the basis of the original measures, shown in the first column. The third quintile is the Median quintile. The second column, TSA, indicates an average of the estimates for the funds in that quintile with stock fixed effects in the panel regression and using the Hjalmarsson (2010) method. For the GT measure, this is followed by the components of TSA broken down with the DGTW benchmark as dynamic selection (DSel) and benchmark timing (Timing). The fifth column in Panel A and the third column in panels B and D is the AAR extracted with the fixed effects. BH is the buy-and-hold drift component of the AAR. The net AAR subtracts the buy-and-hold drift from the AAR. The units are percent per year. HML is the difference between the top and bottom quintile measures. The sample period is from 1980 to 2012, and the number of active funds is 3596.

Table 5: Multiple Comparisons

Models	GT		DGTW		FM	
	Left tail	Right tail	Left tail	Right tail	Left tail	Right tail
Original Measures:						
Critical value	-2.64	2.70	-3.35	5.87	-4.22	5.83
Percent Exceeding	7%	7%	4%	1%	2%	1%
Fractions Nonzero	2%	11%	5%	6%	6%	0%
TSA:						
Critical value	-2.59	2.42	-3.29	3.51	-3.27	3.66
Percent Exceeding	15%	7%	6%	2%	6%	2%
Fractions Nonzero	5%	9%	32%	2%	33%	2%
Net AAR:						
Critical value	-2.12	1.26	-5.15	4.05	-5.00	4.11
Percent Exceeding	**37%**	**53**%	2%	3%	3%	3%
Fractions Nonzero	15%	4%	15%	11%	26%	9%
TSA + net AAR:						
Critical value	-2.60	2.18	-5.72	4.79	-5.88	4.60
Percent Exceeding	**19%**	10%	1%	1%	1%	1%
Fractions Nonzero	9%	9%	30%	3%	40%	1%

This table imposes the null hypothesis that TSA is zero by randomly scrambling the returns for a given fund's stocks through time, using the original porfolio weights to measure the performance. To remove the AAR relation between average weights and stocks' alphas, we randomly scramble the returns for a given fund-month across the stocks held by the fund that month, using the original portfolio weights to measure the performance. Each experiment generates a critical value for the performance measure, such that 5% of the estimates under the null exceed the critical value. This table shows the critical values and the fractions of funds in the original data that exceed the critical values. The standard error of these fractions is about 4.4%. The rows labelled Fractions Nonzero present estimates of the fractions of funds with positive and negative performance using false discovery methods following Barras, Scaillet and Wermers (2010).

Table 6: Fund flows and Performance Measures

Models	GT		DGTW		FM		FM-CAPM	
Controls?	N	Y	N	Y	N	Y	N	Y
Original Measures:								
Slope $\beta_{F\varepsilon}$	8.61	8.25	16.02	15.56	13.27	11.78	10.28	9.69
t-ratio	**9.63**	**8.47**	**12.01**	**11.59**	**8.05**	**7.72**	**4.52**	**4.59**
Frac Correct	54.31	54.12	58.01	57.78	56.63	55.89	55.14	54.85
TSA:								
Slope $\beta_{F\varepsilon}$	1.48	1.71	3.63	3.02	4.68	2.86	2.91	1.79
t-ratio	0.81	1.29	**2.33**	**2.43**	**2.04**	**2.01**	1.35	1.33
Frac Correct	50.74	50.86	51.81	51.51	52.34	51.43	51.45	50.89
AAR:								
Slope $\beta_{F\varepsilon}$	4.58	4.60	7.32	6.96	3.23	1.94	1.08	0.22
t-ratio	**4.75**	**5.74**	**11.02**	**13.11**	1.24	1.06	0.40	0.11
Frac Correct	52.49	52.30	53.66	53.48	51.61	50.97	50.54	50.11
Net AAR:								
Slope $\beta_{F\varepsilon}$	4.63	4.53	5.45	4.70	6.10	5.13	5.82	5.09
t-ratio	**6.03**	**6.67**	**6.65**	**7.08**	**4.77**	**5.73**	**5.22**	**6.43**
Frac Correct	52.32	52.26	52.72	52.35	53.05	52.26	52.91	52.54
BH Drift:								
Slope $\beta_{F\varepsilon}$	0.79	0.91	0.70	1.42	0.06	-0.82	-0.93	-1.66
t-ratio	0.66	0.99	0.80	1.85	0.02	-0.44	-0.36	-0.90
Frac Correct	50.39	50.45	50.35	50.71	50.03	49.59	49.53	49.17

The sign of new money flows is regressed in a panel on the sign of abnormal performance for various models and performance components. The coefficient on performance is β_F . The original measures are calculated from the "up-to-t" version of the Haj10 estimator and use data to month t-13 relative to the flows. The models and their components are defined in the previous tables. Frac Correct is $(100 + \beta_{F\varepsilon})/2$, and indicates the probability that the signs of the flows and performance match.

Table 7: Controlling for Dividend Yield

Models	Low D/P		Median		High D/P		HML	
D/P Controls?	N	Y	N	Y	N	Y	N	Y
GT:								
Classical measure	0.13	0.12	0.09	0.06	0.16	0.05	0.03	-0.07
t-ratio	**2.17**	**2.91**	**2.29**	**2.47**	**3.82**	1.53	0.64	**-2.71**
TSA:								
Slope β	0.12	0.12	0.07	0.07	0.10	0.04	-0.02	-0.08
t-ratio	**2.40**	**2.68**	1.86	**2.34**	**2.75**	1.67	-0.67	**-2.90**
DGTW:								
Classical Measure	-0.05	-0.13	-0.01	-0.14	0.17	-0.09	0.22	0.04
t-ratio	-1.36	**-4.67**	-0.46	**-8.37**	**8.27**	**-8.49**	**5.06**	1.28
TSA:								
Slope β	-0.02	0.03	-0.03	-0.00	0.03	-0.01	0.05	-0.04
t-ratio	-0.54	0.94	-1.01	-0.14	1.17	-0.70	1.66	**-2.17**
FM:								
Classical Measure	-0.06	-0.09	0.05	-0.07	0.36	0.02	0.42	0.11
t-ratio	-0.20	-0.56	0.18	-0.53	1.41	0.20	**3.80**	1.72
TSA:								
Slope β	-0.10	0.02	-0.07	0.00	0.05	-0.00	0.15	-0.02
t-ratio	-0.95	0.31	-1.03	0.00	0.77	-0.02	**2.21**	-0.64

Holdings-based performance measures are estimated with or without a control variable in the panel regression. The control variable is the dividend/price ratio (D/P) of a stock, defined using the previous 12-months dividends per share. The coefficient on the lagged portfolio weight is shown with its t-ratio. Funds are sorted into quintiles on the basis of their fund level average D/P ratios and the averages for selected quintiles are show. HML is the high-low quintile spread. GT is the Grinblatt and Titman (1993) measure, DGTW is Daniel, Grinblatt, Titman and Wermers (1997) and FM is the Ferson and Mo (2016) measure.

Table 8: Persistence

Panel A: Predicting DGWT performance

		AAR		TSA		Original Measure	
		HML	T	HML	T	HML	T
	GT	1.56	(3.23)	-0.08	(-0.26)	1.63	(3.08)
1Q	DGTW	1.70	(2.37)	-0.17	(-0.66)	1.87	(2.16)
	FM	1.42	(2.04)	-0.23	(-0.71)	1.61	(1.91)
	GT	0.92	(2.26)	-0.02	(-0.06)	0.85	(2.01)
1y	DGTW	0.82	(1.60)	-0.03	(-0.15)	0.81	(1.48)
	FM	0.60	(1.18)	-0.16	(-0.63)	0.69	(1.29)
	GT	0.38	(1.53)	0.03	(0.11)	0.29	(1.03)
2y	DGTW	0.18	(0.69)	0.05	(0.26)	0.06	(0.18)
	FM	0.19	(0.78)	-0.19	(-0.86)	0.02	(0.06)
	GT	0.29	(1.74)	0.20	(1.11)	0.30	(1.64)
5y	DGTW	0.35	(2.19)	0.27	(1.94)	0.34	(1.23)
	FM	0.14	(1.15)	0.12	(0.96)	0.12	(0.52)

Panel B: Predicting BvB Value Added

	GT	0.77	(1.93)	-0.14	(-0.48)	0.82	(1.83)
1Q	DGTW	0.73	(1.56)	-0.14	(-0.51)	0.59	(1.11)
	FM	0.73	(1.55)	-0.10	(-0.33)	0.76	(1.30)
	GT	0.20	(0.54)	-0.19	(-0.68)	0.21	(0.56)
1y	DGTW	0.16	(0.41)	-0.37	(-1.67)	0.27	(0.58)
	FM	0.06	(0.15)	-0.26	(-1.13)	0.39	(0.72)
	GT	-0.05	(-0.21)	-0.24	(-0.93)	-0.11	(-0.46)
2y	DGTW	-0.02	(-0.08)	-0.37	(-1.23)	-0.03	(-0.15)
	FM	-0.11	(-0.64)	-0.38	(-1.28)	0.01	(0.04)
	GT	0.06	(0.51)	0.16	(1.20)	0.02	(0.11)
5y	DGTW	0.06	(0.40)	0.08	(0.61)	0.14	(0.84)
	FM	-0.05	(-0.34)	0.04	(0.28)	0.19	(0.77)

This table sorts funds into deciles on their up-to-t versions of the performance measures, their AAR and their TSA components. Future DGTW performance is examined in Panel A and future Value Added, which is DGTW alpha multiplied by the total net assets. The future periods indicated in the rows. GT is the Grinblatt and Titman measure, DGTW is the Daniel, Grinblatt, Titman and Wermers measure and FM is the Ferson and Mo measure. The HML is the difference between the DGTW alphas of the top and bottom decile, and the t-ratios are computed from the monthly return differences as in Fama MacBeth (1973), adjusted for serial dependence.