HOW MUCH DOES SIZE ERODE MUTUAL FUND PERFORMANCE? A REGRESSION DISCONTINUITY APPROACH

Jonathan Reuter
Eric Zitzewitz

Working Paper 16329
http://www.nber.org/papers/w16329

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

We thank Jonathan Berk, Joe Chen, Edie Hotchkiss, Jeffrey Pontiff, Phil Strahan, and participants at the 2009 Financial Research Association annual meeting for helpful comments. Any remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Much Does Size Erode Mutual Fund Performance? A Regression Discontinuity Approach
Jonathan Reuter and Eric Zitzewitz
NBER Working Paper No. 16329
September 2010, Revised December 2011
JEL No. G14,G23,G24

ABSTRACT

Although mutual funds exhibit little ability to persistently outperform their peers, money flows into funds with the highest past returns. Berk and Green (2004) rationalize these patterns by arguing that more-skilled managers manage more assets but, because of diseconomies of scale, generate the same expected returns as less-skilled managers. To identify the causal impact of fund size on performance, we exploit the fact that small differences in mutual fund returns can cause discrete changes in Morningstar ratings that, in turn, generate discrete differences in mutual fund size. Our regression discontinuity estimates yield little evidence that fund size erodes fund returns.

Jonathan Reuter
Carroll School of Management
Boston College
224B Fulton
140 Commonwealth Avenue
Chestnut Hill, MA 02467
and NBER
reuterj@bc.edu

Eric Zitzewitz
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
eric.zitzewitz@dartmouth.edu

The mutual fund literature is responsible for two well-known stylized facts. The first fact, based on more than thirty years of research, is that actively managed mutual fund returns exhibit little ability to persistently outperform their peers (e.g., Jensen (1968) and Carhart (1997)). The second fact is that new money flows disproportionately into those actively managed funds with the highest past returns (e.g., Chevalier and Ellison (1997) and Sirri and Tufano (1998)). The traditional academic interpretations of these facts are that fund managers are unskilled and fund investors are unsophisticated.

Berk and Green (2004) challenge these interpretations. They argue that both facts are consistent with a model that combines skilled managers with diseconomies of scale in asset management. In their model, rational investors chase performance to the point that expected future returns are equalized across funds. In equilibrium, more-skilled managers manage more assets but—precisely because of the diseconomies of scale associated with managing more assets—earn the same expected future return as their less-skilled peers. Berk (2005) goes further, arguing that the traditional interpretations of these stylized facts are "myths" and that the Berk-Green model shows that "most active managers are skilled." Berk and Green's interpretations have quite different implications for our view of financial markets (i.e., they are easier to beat than we thought) and investors (i.e., they are harder to fool than we thought), which, in turn, have important implications for public policy, and for the evaluation of fund managers. However, the empirical relevance of the Berk-Green model depends crucially on the degree of scale diseconomies in asset management.

Our goal in this paper is to measure the causal impact of fund size on fund performance. To motivate our empirical strategy, it is helpful to view the existing evidence through the lens of Berk and Green's (2004) model. In a study that is both representative and widely cited, Chen, Hong, Huang, and Kubik (2004, hereafter CHHK) regress mutual fund returns on lagged fund size and other observable characteristics. They find that a fund that is a log order of magnitude larger earns risk-adjusted returns that are 2 to 3 basis points per month lower.[1] If we were to interpret this difference as the causal effect of fund size on returns, we would conclude that diseconomies could not be masking a meaningful amount of performance persistence. First, we know that a fund that outperforms its peers by one log percentage point this year will be 2-5 log percentage points larger next year (one log percentage point from returns mechanically increasing assets, and the other 1-4 log percentage points from the flow-performance relation).[2] Second, CHHK's estimate implies that a fund that is one log percentage point larger will earn returns that are about 0.003 log percentage points lower over the next 12 months. Combining these two estimates implies that a fund that outperforms its peers by one percentage point this year will suffer a 0.6-1.5 basis point penalty next year. In other words, if we interpret CHHK's estimate as an estimate of the causal effect of fund size on performance, the effect described in Berk and Green will cause us to underestimate an annual AR(1) coefficient by 0.006-0.015. Given that we estimate the annual AR(1) coefficient to be approximately 0.1, the estimated diseconomies of scale in CHHK are too small to meaningfully affect our views about the level of return persistence.

However, it is important to note that the calculation above is not an appropriate test of the Berk-Green model. If fund size is endogenously related to expected future returns, in equilibrium, fund size will be uncorrelated with future returns, thereby frustrating standard approaches to estimate diseconomies of scale. Even if we allow for the possibility that fund sizes are out of equilibrium, the estimates in CHHK (and other studies) will underestimate the actual diseconomies

[1] Chen, Hong, and Kubik (2008) and Massa, Reuter, and Zitzewitz (2010) estimate similar partial correlations between fund size and fund returns, although neither paper is focused on the relation between fund size and returns.

[2] We take our range from the graphs of the inflow-performance relationship for the "young" (<2 years) and "old" (>10 years) funds in Chevalier and Ellison (1997), but these slopes have been replicated in many other studies.

of scale if larger funds have more-skilled managers.[3]

Indeed, Berk-Green effects may analogously inform the interpretation of any study involving fund, or even trade, size. For example, Edelen, Evans, and Kadlec (2007) and Yan (2008) provide evidence that trading costs are higher in larger funds, likely depressing returns. In a Berk-Green model, larger funds have more skilled managers, so if skill mitigates trading costs, the partial correlation should yield a conservative estimate of the causal effect of size in increasing costs. A recent literature (summarized in Bessembinder and Maxwell, 2008) finds that corporate and municipal bonds trading costs decline sharply with trade size. While this relationship may largely reflect a causal, "economy of scale," relationship, the average client in a large trade is likely more skilled at search and negotiation in dealer markets, and this may contribute to the size-cost correlation. Consistent with this possibility, Zitzewitz (2011) provides evidence that clients in large corporate bond trades are more likely to be insurance companies and that insurance companies trade at lower costs than the average client, controlling for trade size. Finally, Gutierrez, Maxwell, and Xu (2009) find no evidence of a size-return correlation for bond funds, and suggest that the difference with the results for stock funds reflects differences in economies of scale in trading costs for stocks and bonds. An alternative, Berk-Green hypotheses might be that the size-skill relationship is stronger for bond funds, where skill may be more important or more readily inferred from returns.

To identify diseconomies of scale in asset management, separately from the effects of other factors that covary with size, we require a natural experiment—something that causes an increase in fund size for reasons that are related

[3] Controlling for additional fund characteristics, as most studies comparing large and small funds do, does not change the fundamental prediction that the partial correlation between fund size and expected returns should be zero, even in the presence of scale diseconomies. When observable fund characteristics impact expected returns, investors should allocate dollars across funds such that expected returns are equal conditional on those characteristics.

to future returns only through diseconomies of scale. We identify such an experiment using a regression discontinuity approach. Our insight is that small changes in fund returns can have discontinuous impacts on fund flows through their impact on the fund's Morningstar rating. For example, as a fund's within-category Morningstar performance ranking increases from the 89th percentile to the 90th percentile, it's Morningstar rating increases from four stars to five stars. Under the assumption that manager skill varies continuously across each of the Morningstar rating thresholds, we can use high frequency data on Morningstar performance rankings to identify the casual impact of Morningstar rating thresholds on fund inflows. Then, because this source of fund inflows is uncorrelated with manager skill (and other factors affecting future returns), we can use these inflows to identify the causal impact of fund size on fund performance. In other words, we are using small deviations from the rational behavior assumed in the Berk-Green model to measure the extent of diseconomies of scale.

We have four main empirical findings, based on monthly data from Morningstar that covers virtually every mutual fund in operation between December 1996 and August 2009. First, in our first-stage regressions, we show that mutual funds just above the threshold for a Morningstar rating receive incremental net flows over the next six months that are equal to approximately 2.5 percent of assets under management. Second, looking out over the next 6-24 months, we find little evidence of diseconomies of scale. Our reduced-form estimates of the impact of incremental net flows on returns are largely *positive* during the first six months and largely negative during the subsequent eighteen months, but *none* of the estimates are statistically different from zero. In other words, within the full sample of funds, the exogenous variation in fund size that we exploit has little impact on fund returns.

Third, when we shift our focus to subsamples based on investment objectives (e.g., small-cap equity funds or sector funds), we continue to find little evidence of diseconomies of scale. For example, despite the fact that incremental inflows into sector funds reach 14.0 percentage points by month 24, and despite the fact that sector funds should be a good category in which to test for diseconomies of scale, we cannot reject the hypothesis that the diseconomies implied by our first-stage and reduced-form regressions equal those obtained via standard OLS regressions. The only subsamples for which we can reject the hypothesis that the IV estimate equals the OLS estimate are mid-cap equity funds and municipal bond funds.[4] And, within both of those subsamples, our IV estimates imply (small) positive economies of scale.

Finally, we adjust standard OLS estimates of performance persistence for potential diseconomies of scale. Because we find little evidence of diseconomies of scale, our median corrected AR(1) estimate of 0.088 is virtually identical to our uncorrected estimate of 0.090. Moreover, the upper bound of the 95% confidence interval for the corrected persistence coefficient is 0.137 in the full sample of funds, and even lower in some subsamples. Based on our regression discontinuity estimates, it is hard to attribute the well-known lack of performance persistence to diseconomies of scale.

The remainder of our paper is organized as follows. In Section I, we describe the process that Morningstar uses to determine ratings, as well as our data. In Section II, we outline our empirical strategy and discuss our identifying assumption. In Section III, we show that share classes (and funds) with return patterns that place them just above a Morningstar ratings threshold receive higher flows than share classes (and funds) with return patterns that place them just below the same Morningstar ratings threshold. In Section IV, we use the findings

[4] One might argue that when OLS and IV estimates are equal, there is no scientific contribution from having done the IV. We disagree, particularly when there is a plausible alternative explanation for the OLS result that the IV approach rules out. Two famous examples come from the literatures on smoking and lung cancer (e.g., Doll, 1998) and wages and education (e.g., Card, 1999). In both cases, initial correlative evidence was dismissed on the grounds that there were very plausible confounding factors, and it was left to later work to establish a causal relationship.

from Section III to test for diseconomies of scale, both overall and within samples of funds focused on specific asset classes. In Section V, we adjust estimates of return persistence for diseconomies of scale. In Section VI, we conclude.

I. Morningstar Ratings and Fund Characteristics

Our identification strategy relies on the discrete nature of Morningstar ratings. It also relies on the fact that, because Morningstar ratings are based on past returns, we can identify funds near rating thresholds. In this section, we describe how Morningstar ratings are determined. We then describe our sample.

A. Morningstar Ratings

Morningstar rates mutual fund share classes on a scale that ranges from one star (the lowest possible rating) to five stars (the highest possible rating). The rating assigned to each mutual fund share class depends on its relative performance within its Morningstar-determined investment category over the prior 3 years, 5 years, and 10 years, "after adjusting for risk and accounting for all sales charges."[5] Morningstar does not rate mutual fund share classes that are less than three years old.

For mutual fund share classes between the age of three and five years, the Morningstar rating depends entirely on its relative performance over the prior 36 months. "Within each Morningstar Category, the top 10% of funds receive five stars, the next 22.5% four stars, the middle 35% three stars, the next 22.5% two stars, and the bottom 10% receive one star."[6] Therefore, small differences in past

---

[5] Morningstar changed various detailed of its ratings process in June 2002. See Blume (1998) for a description of the rating system used from 1996-2002 and http://quicktake.morningstar.com/DataDefs/FundRatingsAndRisk.html for Morningstar's description of their current ratings process. The most significant change was that the number of Morningstar Categories increased from four on May 2002 (Domestic Equity, International Equity, Taxable Bonds, and Munipal Bonds) to 48 on June 2002, eventually growing to 81 in August 2009. The new Morningstar Categories better reflect actual investment styles (e.g., distinguishing domestic equity funds that focus on large-cap growth from those that focus on small-cap value). Morningstar also changed the method used to risk-adjusting returns, and made the relative importance of 5 and 10-year returns depend on whether a fund had experienced style drift.

[6] See http://quicktake.morningstar.com/DataDefs/FundRatingsAndRisk.html.

returns, such as going from the 10th percentile to the 11th percentile, or from the 89th percentile to the 90th percentile, result in discrete changes in Morningstar ratings. These discrete changes are evident in Figure 1, in which we plot Morningstar ratings for all share classes that are less than 5 years old against Morningstar's risk-adjusted, within category return percentile. Figure 1 also provides graphical evidence that (residual) flows increase sharply around ratings thresholds.[7] (We present more formal evidence in Section III.)

For share classes between the age of 5 and 10, Morningstar determines separate ratings based on the prior 36 months and the prior 60 months, and "averages" the underlying ratings to calculate an overall integer rating. In Figure 2, we show how relative performance over the prior 36 and 60 months maps into a share classes' overall rating. The pattern reveals that Morningstar calculates a fund's overall rating as a 60-40 average of the 5-year and 3-year integer ratings, causing it to "round up" when the better performance is over the longer horizon.[8] For example, a share class with a 36-month return that puts it at the 89th percentile (four stars) and a 60-month return that puts it at the 90th percentile (five stars), receives an overall rating of five stars. In contrast, a share class with a 36-month return that puts it at the 90th percentile (five stars) and a 60-month return that puts it at the 89th percentile (four stars), receives an overall rating of four stars. To the extent that we're willing to assume that the managers of these two funds are similarly skilled (conditional on current assets under management), and that the five-star fund receives higher residual flows, we can study the impact of these incremental flows on future returns.

[7] The residual flows in Figure 1 come from versions of the baseline flow regression in Section III that omit the Morningstar within-category percentile ranking and discontinuity dummy variable.

[8] After June 2002, Morningstar began giving older history less weight when funds had experienced style drift. To make Figure 2 more transparent, we exclude these funds from the picture. Depending on how much style drift was experienced and when it was experienced, a fund's 3-year history can receive more than 50 percent of the weight, causing the rounding to occur in the other direction.

While the staircase boundaries between overall ratings may strike readers as an unusual methodological choice by Morningstar, it is helpful from the perspective of our research, since this approach increases the number of funds that are very close to a rating boundary. For share classes that are more than 10 years old, Morningstar's overall rating depends on the average of the 3-year, 5-year, and 10-year ratings. For these share classes, thresholds between ratings are conceptually similar to those in Figure 2. However, because these thresholds relate to three underlying ratings, they must be plotted in three dimensions.

B. Sample Construction

To study the impact of mutual fund flows on mutual fund returns, we obtain data from Morningstar Principia CDs. Our sample consists of all open-end mutual funds that have at least one share class rated by Morningstar. Because Morningstar does not rate share classes that are less than three years old, mutual funds enter our sample when their oldest share class reaches three years of age. The fact that we only study funds in the time period in which they appear on a Morningstar CD limits the influence of incubation bias (Evans (2010)) on our results. While incubation bias might help to explain why funds appearing on a Morningstar CD for the first time have average Morningstar ratings about a quarter point above older funds, our analysis of future inflows and performance uses only non-backfilled data. Consequently, our estimates of scale diseconomies should be unaffected by incubation bias.

Our data begin in December 1996 and end in August 2009.[9] Because mutual fund share classes can earn different Morningstar ratings and experience different inflows, the unit of observation in our initial analysis of inflows is the share class. As any scale diseconomies would occur at the fund (portfolio) level, how-

[9] We have been unable to obtain data for 12 of the 36 months between January 1997 and December 1999. The missing months are January 1997, February 1997, April 1997, May 1997, July 1997, August 1997, October 1997, November 1997, January 1998, July 1998, January 1999, and November 1999.

ever, in most of our analysis we aggregate variables to the fund level, weighting each share class in proportion to its assets under management in the prior month. In practice, the exact approach we take to weighting share classes has little influence on the results because the average fund gets 84 percent of its assets from its largest share class.

Finally, because Morningstar within-category percentile rankings do not distinguish between actively and passively managed mutual funds, we include the share classes of index funds in our sample when calculating within-category percentile rankings. However, we exclude index funds from all inflow and return regressions. [10]

C. Summary Statistics

In Table 1, we report fund-level summary statistics for the full sample of 491,863 fund-month observations. We also use asset-weighted average Morningstar ratings to assign fund-level ratings, and report summary statistics for each fund-level rating category. Looking across these categories, we see that funds with higher ratings tend to be larger and come from larger families. Funds with higher ratings also tend to charge lower average fees (both in month t and month t+12), tend to offer fewer share classes, and are less likely to charge a sales load. Of course, differences in fees and sales loads follow, at least in part, from the fact that Morningstar ratings are based on returns measured net of fees and loads.

The most interesting differences between funds with higher and lower ratings involve future flows and future returns. Consistent with investors responding to Morningstar ratings (or to the return histories underlying them), we find that funds with higher ratings receive higher net flows over the next 24 months. Rela-

[10] Conceptually, because index funds should not experience diseconomies of scale, including them in our reduced-form regressions could cause us to underestimate the extent of diseconomies of scale. Practically, however, few index funds would have been included in our reduced-form regressions because the vast majority of index funds are rated three stars, with returns well away from thresholds for two or four stars.

tive to other funds in their Morningstar category, the typical five-star funds grows by 23 log percentage points over this period, while the typical one-star funds shrinks by 18 log percentage points. The results presented later imply that of this 41 log percentage point difference, about 9 log percentage point represents a causal effect of the difference in Morningstar ratings on flows, with the remainder being due to investors responding directly to observable fund characteristics included in Morningstar's ratings (e.g., past returns, risk, and loads), other observable characteristics correlated with the ratings (e.g., low expenses), or unobservable characteristics correlated with the ratings (e.g., marketing efforts).[11]

Consistent with prior work on the predictive power of Morningstar ratings (e.g., Blake and Morey (2009)), we find that one-star funds underperform other funds over the next 24 months, but find little difference in the future performance of other funds. The fact that 5-star and 2-star funds perform approximately as well in the future despite 5-star funds experiencing greater inflows does not necessarily imply the absence of scale diseconomies, however. In the Berk-Green model, the 5-star funds attract more inflows because they have more skilled managers, and this skill allows the funds to match the 2-star funds' returns despite managing more assets. For a test for scale diseconomies to be valid in the Berk-Green model, it needs to exploit a source of variation in inflows that is not caused by or correlated with manager skill. Fortunately, the discontinuities in the Morningstar ranking function generate this type of variation.

II. Overview of RD and our Identification Strategy

In order to measure the causal impact of fund size on fund performance, we must identify variation in fund size that is uncorrelated with manager skill. In markets

[11] Prior work examining the relationship between fund inflows and Morningstar ratings uses observable variables to control for these factors. For example, when they include fund fixed effects, Del Guerico and Tkac (2008) continue to find a positive association between stars and flows. As a robustness check, in Section IV.D., we use changes in Morningstar ratings to estimate the extent of scale diseconomies.

with perfectly rational, informed investors, this variation should be impossible to come by. We use a regression discontinuity approach that exploits the fact that mutual funds with past returns immediately above a Morningstar rating threshold receive a discretely higher rating than mutual funds with past returns immediately below the threshold. To the extent that investors place positive weight on Morningstar ratings, funds with risk-adjusted returns immediately above a ratings threshold are likely to receive significantly more inflows than funds with risk-adjusted returns immediately below the threshold.[12]

Our analysis proceeds in two stages. In the first-stage regressions, we estimate the impact of rating thresholds on future flows. Then, we use reduced-form regressions to estimate the impact of rating thresholds on future returns. The identifying assumption is that while inflow will vary sharply at each threshold, the other fund characteristics that might be related to future returns will vary continuously.[13] Under this assumption, our first-stage and reduced-form estimates allows us to measure the extent of diseconomies of scale.

More formally, our analysis focuses on actively managed mutual funds just above and below each rating threshold. For example, with respect to the threshold between four stars and five stars, our first-stage regression predicts log net flows as function of the within-category percentile ranking used to determine

[12] In the Berk-Green model, investors use risk-adjusted past returns to directly infer manager skill. Because perfectly informed investors will not place any weight on Morningstar ratings, flows will vary continuously across Morningstar rating thresholds. To motivate our RD approach, consider a version of the Berk-Green model where many investors observe Morningstar stars rather than risk-adjusted returns, and make inferences about manager skill based on the average characteristics of funds with that rating. In this version of the model, average performance will be equalized across the different Morningstar ratings. However, there will be flow discontinuities at rating boundaries and—because funds just over a boundary will have similar managerial skill to those just under it—the incremental flows will cause funds just over a boundary to underperform. Of course, given this underperformance, savvy investors who directly observe returns will rationally choose to invest in funds just below ratings boundaries. Therefore, for the flow discontinuities that we observe in the data to exist, the number of savvy investors must be limited.

[13] Imbens and Lemieux (2008) and Lee and Lemieux (2009) provide excellent overviews of the regression discontinuity approach.

Morningstar ratings, a dummy variable that indicates whether the within-category percentile ranking the share class *i* of fund *j* in month *t* is above the five star rating threshold, and controls, including multiple controls for past performance and past flows.

$$\text{Flow}_{i,j,t+1} = \delta_{1\text{st}}\text{threshold}_{i,j,t} + \lambda_{1\text{st}}\text{ranking}_{i,j,t} + \beta_{1\text{st}}\text{X}_{i,j,t} + \eta_{i,j,t} \qquad (1)$$

where δ_{1st} measures the discontinuous flow effect associated with the ratings threshold.[14]

In many RD settings, the "forcing variable", which determines whether an observation is above or below the threshold, is exogenous.[15] In our setting, the within-category percentile ranking is not exogenous. However, our identifying assumption is that, because all managers are trying to maximize relative performance, manager skill will vary continuously across the threshold for a higher rating. In other words, while we allow for the possibility that managers with slightly higher returns are slightly more skilled, our identification strategy assumes that skill does not jump in a discontinuous way at the threshold between ratings. The fact that thresholds for different Morningstar ratings depend on within-category performance rankings over as many as three investment horizons increases our confidence that the distribution of manager skill is smoother than the distribution of Morningstar ratings.[16]

---

[14] Following the advice in Imbens and Lemieux (2008), we experimented with more flexible approaches to controlling for the ranking variable, but found that the results varied little from the local linear approach.

[15] For example, to study the impact of the Sarbanes-Oxley Act on firms costs and earnings, Iliev (2010) exploits the fact that U.S. firms with a public float below $75 million in 2002, 2003, or 2004 were allowed to delay compliance with Section 404 until well after the November 2004 date on which slightly larger firms were required to comply.

[16] An common concern in regression discontinuity studies is manipulation of the forcing variable. In our setting,. our identification approach would be threatened if funds with more skilled managers were also able to manipulate their returns in order to place just above a Morningstar cutoff. We conduct several tests for and find no evidence of such manipulation. We test for discontinuities in the density of risk-adjusted returns around Morningstar cuttoffs (McCrary (2008)) and for month-to-month persistence in the discontinuity variable (controlling for the forcing variable).

To estimate fund-level flows, we focus on the discontinuity measure for the fund's largest share class. Then, we estimate a reduced-form regression

$$\text{Return}_{i,j,t+1} = \delta_{\text{rf}}\text{threshold}_{i,j,t} + \lambda_{\text{rf}}\text{ranking}_{i,j,t} + \beta_{\text{rf}}\text{X}_{i,j,t} + \eta_{i,j,t} \qquad (2)$$

where δ_{rf} measures the causal effect of ratings thresholds on returns. Under the assumption that the causal effect of ratings thresholds on flows is unrelated to differences in manager skill, δ_{rf} will capture any diseconomies of scale associated with these flows. Finally, we can estimate the causal impact of flows on returns as the ratio of δ_{rf} to δ_{1st}. The more negative this IV-style estimate, the larger the implied diseconomies of scale.

III. Impact of Morningstar Ratings on Flows

In this section, we present evidence that Morningstar ratings have a causal impact on investor flows. Because our identification strategy exploits the discreteness of Morningstar ratings, and because different share classes of the same mutual fund can receive different Morningstar ratings, we begin by studying the impact of Morningstar ratings on net flows at the share class level. Consistent with equation (1), our general approach is to regress log net flows of share class *i* in month *t+1* on its Morningstar percentile ranking in month *t*, which is our local linear control, and a dummy variable that indicates whether share class *i* is above the threshold for a particular rating in month *t*. Under the assumption that manager skill varies continuously across the rating threshold, the dummy variable will capture incremental flows into the higher-rated fund that are uncorrelated with manager skill.

To quantify these discontinuous flow effects, in Table 2, we estimate separate regressions for each rating thresholds (i.e. one star versus two stars, …, four stars versus five stars), and a pooled regression that combines all four thresholds. In each case, the sample is restricted to those share classes that are within five

---

We also test for discontinuities in lagged flows and returns and the control variables. These results are presented in an Appendix.

percentiles of a rating threshold.[17] For example, when we focus on the threshold between four and five stars, we restrict the sample to share classes with Morningstar rankings between the 85th and 95th percentiles. We further restrict the sample to actively managed funds by excluding any fund that Morningstar identifies as an index fund.

In addition to the variables that we report in Table 2, "Baseline" regressions control for the lagged log size of the share class, portfolio, and family, portfolio turnover, expense ratio, and the presence of loads (front, deferred, and trailing).[18] Because our sample includes the full range of Morningstar categories (i.e., large-cap equity, sector funds, corporate bond funds, etc.), we include a separate fixed effect for each Morningstar category each month. This allows us to compare funds to their peers. In the regressions with "Additional Controls", we supplement the Morningstar percentile ranking variable with controls for Morningstar's measure of risk-adjusted returns, lagged log returns from *t-12* to *t-1*, *t-24* to *t-13*, and *t-36* to *t-25*, and lagged log inflows from *t-12* to *t-1* and *t-3* to *t-1*. Because mutual funds with multiple share classes can appear multiple times in the same month, we cluster standard errors on the fund.[19]

The estimated coefficients on the discontinuity dummy variable are positive and statistically significant for each of the four ratings thresholds, and for the pooled regression that includes all four ratings thresholds. In the baseline regres-

---

[17] In the Appendix, we present robustness checks that vary this five percentile bandwidth.

[18] Edelen (1999) finds that investor flow volatility is associated with lower fund returns and higher trading activity. The incremental net flow caused by Morningstar ratings may be accompanied by higher gross flow that stimulates trading. While we lack Edelen's hand-collected data on gross flows, we do not find evidence of discontinuities in future portfolio turnover at Morningstar boundaries, and including portfolio turnover as a control does not materially affect our results.

[19] Given the large number of regressions estimated in the study and the large number of category-time fixed effects included in each model, it is not practical to cluster standard errors on both fund and month (e.g., following Petersen (2009)). However, when we experimented with two-way clustering on fund and month, we found that the standard errors were quite similar to those reported in the paper. This is likely because all of our regressions include time fixed effects and because the distribution of Morningstar ratings is stable across time periods. We also found that our results were robust to clustering standard errors by family instead of by fund.

sions, the estimates range from 0.337 log percentage points at the boundary between 1 star and 2 stars (significant at the 5-percent level) to 0.946 log percentage points at the boundary between 4 stars and 5 stars (significant at the 1-percent level). When we include additional controls for past returns and past flows, the estimated coefficients decline, but only slightly. For example, within the stacked regression, the estimated coefficient falls from 0.518 to 0.432 log percentage points, but remains statistically significant at the 1-percent level. In other words, share classes that are just above the Morningstar ratings threshold *this month* receive an additional 0.432 log percentage points in net flow *next month*, compared to share classes that are just below the threshold.

If a share class were to maintain its Morningstar percentile ranking just above the threshold for an entire year, this would translate into an additional annual net flow of 5.154 log percentage points. However, this persistence would also call into question our identifying assumption that manager skill varies continuously across rating thresholds. In the last two columns of Table 2, we test this assumption by changing the dependent variable from log net flows in month $t+1$ to log net flows in month t. If the discontinuity in flows in month $t+1$ is due to a discontinuity in flow-producing fund characteristics at the start of month t (rather than to the higher Morningstar rating), we should also find discontinuity effects in month t. Importantly for our empirical strategy, when we shift our focus to current-month flows, only five of the ten estimated coefficients on the discontinuity dummy variable are positive, and only one is statistically significant from zero (at the 10-percent level). These results strongly suggest that the discontinuity in flows in month $t+1$ is solely due to the higher Morningstar rating. Overall, the results in Table 2 provide the “first stage” that we need to study the causal impact of flows on performance.

Of course, to test for diseconomies of scale, we need to study the impact of fund-level flows on fund-level performance. In Table 3, we study the impact of Morningstar ratings on log net flows at the fund level. Because many funds have more than one share class, we need a measure of incremental flows that is aggregated across all of fund *j*'s share classes. Most funds have a main share class that contains most of the assets (traditionally the "A" class for load funds and the "Investor" class for no-load funds). Because other share classes have the same return gross of fees and expenses, within-fund differences in returns (and Morningstar percentile rankings) reflect differences in fees and expenses. The Morningstar rating of the largest share class is generally the one marketed to potential investors, as other share classes either have lower ratings due to higher fees (e.g., B, C, and Service share classes) or impose restrictions on who can purchase them (e.g., Institutional share classes). Our approach is to focus on the discontinuity and ranking variables for fund *j*'s largest share class.[20]

The estimated coefficients in the first two columns of Table 3 are qualitatively similar to those in Table 2, with slightly smaller magnitudes because the denominator is fund-level assets rather than share class-level assets. Seven of the ten coefficients are statistically significantly different from zero at conventional levels, with the lack of a discontinuity between one and two stars being the major exception. Importantly, we continue to find little evidence of a discontinuity in current month flows.

Again, the coefficient estimates from the regressions with additional controls are slightly smaller than the baseline estimates, but the differences are never statistically significant. Because we find the strongest evidence of flow discontinuities at the 3/4 star and 4/5 star boundaries, we focus on these boundaries in later tables. Figure 3A provides graphical evidence of the discontinuity in future

[20] As an alternative, we experimented with taking the highest Morningstar rating and ranking variable across all share classes, on the assumption that this would be the rating marketed to investors. We found very similar results.

inflows at each rating threshold.[21] Figure 3B provides graphical evidence of the lack of the discontinuity in current month inflows. In the appendix, we provide evidence of a lack of discontinuities in the control variables, which further supports our identification strategy.

IV. Testing for Diseconomies of Scale

We now use the incremental flows earned by funds with returns just above rating thresholds to test for diseconomies of scale. We begin by estimating first-stage and reduced-form regressions on the full sample of mutual funds over longer investment horizons. Then, because diseconomies of scale may differ across asset classes, we estimate first-stage and reduced-form regressions for different subsamples of mutual funds. Finally, we compare the diseconomies of scale estimates implied by our first-stage and reduced-form regressions to the diseconomies of scale estimates implied by standard OLS regressions.

A. Evidence from the Full Sample of Mutual Funds

In Table 4, we extend the analysis in Table 3 along two dimensions. First, rather than estimating first-stage regressions focused on log net flows in month $t+1$, we estimate first-stage regressions focused on cumulative log net flows over different investment horizons. Our goal is to measure the long-term impact of rating thresholds on fund flows. Second, for each first-stage regression of log net flows on the discontinuity variable (and full set of controls), we estimate a matching reduced-form regression of log net returns on the discontinuity variable (and same set of controls). Given our identification assumption that flows associated with rating thresholds are uncorrelated with manager skill, these flows should only impact fund returns through diseconomies of scale. The reduced-form regressions are intended to measure this impact. The sample is restricted to actively managed funds, and all standard errors are clustered on fund.

---

[21] Residual flows in Figures 3A and 3B are estimated from the baseline specification in Table 3 but omit the Morningstar within-category percentile ranking and discontinuity dummy variables.

When we restrict attention to net flows in *t+1* and net returns in month *t+1*, we find little evidence of diseconomies of scale. Estimated incremental flows are between 0.41 and 0.63 log percentage points, depending on the rating boundaries we study, and statistically significant from zero at the 10-percent level and lower. In contrast, the estimated coefficients on the discontinuity variable in the return regression are economically small, ranging between 0.00 and 0.01 log percentage points, and are not statistically significantly different from zero at conventional levels.

When we focus on cumulative log net flows beyond month *t+1*, we continue to find that Morningstar rating boundaries are associated with significant incremental flows. For example, in the regressions that focus on the 3/4 star boundary, the incremental flows associated with the discontinuity variable (measured in month *t*) are 0.61 log percentage points in month *t+1*, 1.74 log percentage points through month *t+6*, 2.50 log percentage points through month *t+12*, and 3.41 log percentage points through months *t+24*. These estimates imply that while the effect of an extra Morningstar star in the ranking disseminated during month *t+1* is strongest in that month, the effect of the extra star persists beyond the initial month. There are numerous mechanisms that could produce this effect. Investors may make an initial investment in month *t+1* based on the current-month Morningstar rating, and that initial investment decision may affect the placement of subsequent investments. Investors may also make investment decisions based on an accumulation of signals received over several months. Regardless of the mechanism, our findings about the timing of investor reactions to Morningstar are consistent with prior findings on the timing of investor reactions to media mentions or advertising (e.g., Reuter and Zitzewitz (2006)).

When we examine returns after month *t+1*, we continue to find little evidence that the variation in fund size associated with rating thresholds affects future returns. The strongest evidence of scale diseconomies for the regression that stacks all four rating boundaries appears in month *t+21*, where funds with returns just above a rating threshold receive flows totally 2.09 log percentage points of assets and underperform their peers by 9 basis points. The strongest evidence from the regressions that include only the 4/5 star boundary is in month t+15, where incremental flows of 2.27 log percentage points are accompanied by underperformance of 12 basis points. In neither case, however, is the underperformance statistically significant at conventional levels. And, when we focus on only the 3/4 star boundary or on the 3/4 star and 4/5 star boundaries stacked, none of the reduced-form estimates between month t+1 and t+24 is even negative. In other words, exploiting exogenous variation in fund size due to Morningstar rating thresholds, we find little evidence of diseconomies of scale. In Figures 4A and 4B, we graph the contemporaneous and cumulative flow and return effects presented in Table 4 as a function of time.

B. Evidence from Different Investment Categories

Although we find little evidence of diseconomies of scale within the full sample of mutual funds, we might reasonably expect the degree of diseconomies of scale to vary across asset classes. For example, CHHK find their strongest evidence of diseconomies of scale among small-cap equity funds. More generally, we might expect the strongest diseconomies of scale in asset classes with less liquidity (e.g., municipal bond funds) or where the inflows experienced by a typical fund are large relative to the investment options available (e.g., sector funds).

In Table 5, we re-estimate the first-stage and reduced-form regressions in Table 4 for different sets of mutual funds over four different investment horizons. We use the Morningstar category variable to create the following seven non-overlapping subsamples of mutual funds: large-cap equity; mid-cap equity; small-cap equity; sector funds; international equity; taxable bonds; municipal bonds. (We exclude a small set of funds that do not fall into these categories, such as balanced funds, commodities funds, and target-date retirement funds.) We also create an "All equity" sample that combines large-cap equity, mid-cap equity, small-cap equity, sector funds, and international equity. We focus on cumulative log flows and log returns through month *t+6*, *t+12*, *t+18*, and *t+24*.

The estimated flow and return effects in the first column of Table 5 are for all funds, and match those reported in Table 4. The other columns focus on different types of funds. Looking across the seven non-overlapping subsamples, we see that the estimated flow effects are almost always positive, but also that the standard errors tend to be much larger than in the full sample. The evidence that Morningstar rating thresholds impact flows is strongest for sector funds, taxable bond funds, and municipal bond funds. For sector funds, the magnitudes are quite large, ranging from 6.19 log percentage points in month *t+6* to 13.95 log percentage points in month *t+24*. Given our need to focus on exogenous variation in fund size, it is hard to imagine ever finding exogenous variation in fund size beyond 13.95 log percentage points. Flows effects are also statistically significant when we focus on the "All equity" sample, ranging from 2.65 log percentage points in month *t+6* to 3.73 log percentage points in month *t+24*. The only subsample-horizon first-stage estimates that are negative are for small-cap equity in months t+12 through t+24.

Turning to the reduced-form regressions for the seven non-overlapping subsamples, we see that 15 of the 28 estimated coefficients are negative. However, the only negative coefficient that is statistically different from zero at conventional levels is for small-cap equity in month t+18, when the first-stage estimate is also negative. In contrast, of the 13 reduced-form coefficients that are positive, all eight of the estimates for mid-cap equity and municipal bonds are statistically different from zero. The fact that we find strong flow effects for sector funds but no corresponding return effects argues against meaningful diseconomies of scale. Our overall evidence is more consistent with modest economies of scale.

C. A Comparison of IV and OLS Estimates of Diseconomies of Scale

Our regression discontinuity approach allows us to directly estimate the causal impact of rating thresholds on flows and the causal impact of rating thresholds on returns. However, we are ultimately interested in measuring the causal impact of flows on returns. To obtain an IV-style estimate of the diseconomy of scales for a particular subsample of mutual funds and investment horizon, we scale the estimated coefficient from the reduced-form by the estimated coefficient from the first stage. For example, for the "All equity" sample of funds through month *t+24,* one log percentage point in incremental flows is associated with incremental returns that are 0.06 log percentage points lower. (The "IV" estimate of -0.06 equals -0.22 divided by 3.73.)

Table 6 reports IV estimates for different sets of mutual funds and investment horizons alongside the first-stage and reduced-form estimates (from Table 5). Eleven of the 25 IV estimates are negative.[22] Among the seven mutually exclusive categories of funds, nine of the 19 IV estimates are negative. However, the standard errors associated with many of the estimates are quite large, particularly at longer time horizons and in categories with smaller inflow effects.

In the last several columns of Table 6, we compare our IV estimates to the partial correlation between fund size and fund returns that we estimate within the same sample using standard OLS regressions. Specifically, the partial correlation for each asset class and investment horizon is estimated as the coefficient on fund size in a regression of future returns on the variables listed under fund characteristics in Table 1, a control for past-12-month log returns, and a separate fixed effect for each Morningstar category each month. Consistent with Berk and Green's prediction, each of the negative IV estimates is significantly more negative than the corresponding OLS estimate. However, the average IV estimate is 0.028 ver-

[22] The fact that the estimated flow effects are negative for small-cap equity funds for months t+12, t+18, and t+24 prompts us to drop these subsample-horizon combinations from Tables 6 and 7.

sus an average OLS estimate of -0.003. Similarly, the median IV estimate is 0.012 versus a median OLS estimate of -0.003. Because of the larger standard errors on the IV estimates, we can only reject the hypothesis that the OLS and IV estimates are equal in the six cases where the IV estimate is statistically significant from zero (the *p*-values of the Hausman tests range from 0.02 to 0.09). And, in all six of these cases, which correspond to mid-cap equity and municipal bond funds, the IV estimate is actually more positive than the OLS estimate.

D. Robustness: Estimating Diseconomies of Scale From Changes in Morningstar Ratings

To test for diseconomies of scale, we have been comparing the future flows and future returns of funds whose current returns are near thresholds for specific Morningstar ratings. This approach is inherently cross-sectional. The rating thresholds that we use in month t are those that Morningstar uses to determine its rating in month t. In this section, we take a different approach and use time-series changes in fund ratings to test for diseconomies of scale. When we modify our baseline flow regression to include both the Morningstar star rating in month $t-1$ and the change in Morningstar star ratings between month t and $t-1$, the (unreported) estimated coefficient on the change in ratings variable is 4.6% (standard error of 0.2%). This suggests that gaining one star is associated with a 4.6% increase in fund size over the next 12 months. While 4.6% is approximately three times the full-sample estimate of 1.6% in Table 4, time-series changes in fund ratings are less plausibly exogenous than the cross-sectional differences in fund ratings that we exploit above.

To identify exogenous changes in Morningstar ratings, we exploit a significant change in how Morningstar ratings are determined. Through May 2002, fund ratings are based on rankings “across four broad asset classes (U.S. stock, international stock, taxable bond, and municipal bond).” However, beginning in June 2002, funds are “ranked and rated within nearly 50 Morningstar Categories."[23] This change in methodology increases the ratings of the top performing funds in investment categories with lower average returns, and decreases the ratings of the not-top-performing funds in investment categories with higher average returns. The standard deviation of the variable measuring changes in ratings jumps from 0.3 stars within the full sample to 0.9 stars in June 2002. When we limit the flow regression sample to June 2002, the estimated coefficient on the change in ratings variable is again 4.6% (standard error of 1.0%).

The advantage of focusing on June 2002 is that, if we treat the changes in ratings in this month as being purely exogenous, they provide another strong first stage. The disadvantage is that we are limited to a single (noisy) cross-section of mutual fund returns. In the reduced-form regression, the coefficient on the change in ratings variable is 0.19 log percentage points (standard error of 0.16). Combining the first-stage and reduced-form estimates, a one log percentage point increase in fund size is associated with a 0.04 log percentage point *increase* in fund returns (standard error of 0.03). This effect is statistically and economically indistinguishable from the 0.01 log percentage point increase that we find for "All funds" in Table 6. Therefore, regardless of whether we focus on our preferred regression discontinuity estimates or estimates based on changes in Morningstar ratings, we find little evidence that fund size erodes performance.

E. How are Incremental Flows Invested?

There are two reasons why incremental flows might not depress fund returns. First, incremental flows may be closet indexed rather than actively managed. Second, incremental flows may be used to deepen existing positions, but without generating enough price impact to significantly harm returns. For example, for sector funds in month *t+12*, flows rise 8.84 log percentage points and after-fee returns fall 0.14 log percentage points. The IV estimate is that a 1 percentage

[23] See the Morningstar Press Release titled "Morningstar, Inc. to Change "Star Rating" for Funds", dated April 22, 2002.

point increase in size is associated with a 2 basis point decrease in returns (standard error is 6 bp). This decrease is consistent with incremental flows being invested at prices that are approximately 2% too high, which may be plausible given that sector funds are constrained to a limited number of stocks.

In Table 7, we examine how incremental flows are invested. We begin by estimating a reduced-form regression with the number of equity holdings as the dependent variable. This specification measures the casual impact of incremental flows associated with the 4/5-star and 3/4-star boundaries on the number of equity holdings. The sample is limited to non-speciality domestic equity funds for which we observe quarterly holdings. We do not find any evidence that the number of holdings increases in response to incremental flows. Instead, our findings suggest that incremental flows are used to deepen existing positions, which is what Pollet and Wilson (2008) find when they treat net flows as exogenous.[24]

Next, we sort the sample of domestic equity funds into two groups based on the number of stocks they hold. Our thinking is that funds with fewer stocks than the median fund (within the same investment objective in the same month) may be more likely to actively manage their incremental flows. If so, funds with more concentrated portfolios may be the most likely to exhibit diseconomies of scale. Although the standard errors are higher than in our main tables, we find some evidence that incremental flows are higher in funds with more concentrated portfolios. This finding suggests that investors prefer to invest in more actively managed funds. However, we find no evidence that funds with more concentrated portfolios earn lower returns than funds with less concentrated portfolios.

F. Summary

Taken together, our approach tests for scale diseconomies in many places in our

[24] In unreported regressions, we test for changes in the Cremers-Antijisto (2009) active share measure. The estimated coefficients are positive, suggesting that incremental flows are driving portfolio weights further from benchmark weights, but the standard errors are large.

data. The multiple Morningstar rating boundaries allows us to test for scale diseconomies among funds that are experiencing net inflows (at the 4/5-star boundary) and net outflows (at the 2/3-star boundary). We can separately examine asset classes where Morningstar rating effects are large (e.g., sector funds and international equity) or modest but precisely estimated (large-cap equity, municipal bonds). We can also separately examine asset classes where diseconomies of scale are most likely because managers have limited securities to chose from (sector funds), where there is more room to invest additional assets (large-cap equity), or where economies of scale may be most likely, due to large trades having lower trading costs (bonds). The change in rating system in 2002 gives us a completely different, but still plausibly exogenous, source of identification, as well as a few examples of rating changes of more than one star. None of these estimates support the view that diseconomies of scale meaningfully erode performance.[25] Finally, when we examine how incremental flows are invested by domestic equity funds, we find that they are used to deepen existing positions. In liquid financial markets, this practice is unlikely to have a significant impact on returns.

V. Adjusting Performance Persistence for Diseconomies of Scale

Finally, we turn to the important question of whether our estimated diseconomies of scale have an economically significant impact on estimates of performance persistence. Berk and Green (2004) show that the combination of diseconomies of scale with endogenous fund flows will cause researchers to underestimate the true degree of performance persistence in the mutual fund industry. In Table 8, we adjust measures of performance persistence for the causal impact of flows on performance. To begin, we estimate standard OLS regressions that predict future

[25] Our empirical strategy implicitly assumes that incremental flows are being actively managed. However, in the Berk-Green (2004) model, once manager skill has been exhausted, incremental flows will be indexed. In the Appendix, we assume that funds with relatively more concentrated portfolios are more likely to actively manage incremental flows, and test whether our reduced-form estiamtes are larger within this sample. We do not find any evidence that they are.

returns from past 12-month returns, as well as the category-by-month fixed effects and the control variables included in Tables 2-7. We report the estimated coefficient on the past return measure, and its standard error, in the first set of columns. Within the full sample of funds, assuming a 12-month horizon, the estimated coefficient is 0.091. However, this estimate will be downwardly biased if past returns attract incremental flows and there are diseconomies of scale.

We use a similar set of regressions to predict the impact of past returns on log net flows. We report the estimated coefficient on the past return measure, and its standard error, in the second set of columns. For the full sample of funds, an additional log percentage points in returns over the past 12 months is associated with 0.947 log percentage points in additional flow. The larger the diseconomies of scale associated with these additional flows, the greater the downward bias in the persistence coefficient.

Finally, we use the diseconomies of scale that we estimated in the prior section to adjust the persistence coefficient. Specifically, we estimate the "Corrected persistence coefficient" as the "Persistence coefficient" minus the "Flow coefficient" times the "Causal effect of flows" from Table 6.[26] For the full sample of funds and a 12-month horizon, adjusting for diseconomies of scale *decreases* the persistence correlation from 0.091 to 0.079. If we use the standard errors from Table 6 to construct a 95% confidence interval for the corrected persistence coefficient, we find that it ranges from 0.022 to 0.137. We cannot reject the hypothesis that the "Persistence coefficient" and "Corrected persistence coefficient" are equal. However, we can reject at the 1-percent level the hypothesis

[26] The "Persistence coefficient" is the increase in expected next-period log percentage point return associated with a one log percentage point increase in 12-month past returns. The "Flow coefficient" times the "Causal effect of flows" is the expected log percentage point decrease in next period's return based on the expected log percentage point increase in flows times the expected diseconomies of scale associated with the incremental flow. The "Corrected persistence coefficient" removes the impact of return-induced flows on the expected log percentage point increase in next period's return.

that the corrected persistence coefficient is equal to 0.42, which is the value implied by Berk and Green's calibration.[27]

Within investment categories, the corrected persistence coefficient is less precisely estimated. As we possess data on virtually every U.S. mutual fund in operation between December 1996 and August 2009, it may prove difficult to significantly increase the statistical power of our tests. Nevertheless, the upper bound of the 95% confidence interval is always well below 0.42, and tends to be lower in categories where we expect diseconomies of scale to be largest. For sector funds the corrected persistence coefficient is slightly higher than the unadjusted persistence coefficient (0.080 versus 0.069), but the upper bound of the 95% confidence interval is only 0.167. In contrast, we obtain the largest upper bound (0.329) for "Large-cap Equity", a category for which we expect diseconomies of scale to be relatively small. Our inference, based on our empirical strategy for identifying exogenous variation in fund size, is that correcting for scale diseconomies does not significantly affect our view of performance persistence.

VI. Conclusion

The Berk-Green model poses a serious challenge to the common academic view that mutual fund managers are unskilled and mutual fund investors are unsophisticated. The prediction that more skilled managers will manage larger funds also poses a serious challenge to existing evidence on diseconomies of scale. We use a regression discontinuity approach to determine the practical importance of the endogeneity problem implied by the Berk-Green model. Specifically, we use the discrete changes in flows associated with discrete changes in Morningstar ratings to identify flows that should only impact fund returns through diseconomies of

[27] In Berk and Green's calibration exercise, managers' skill, defined as the annualized alpha they would achieve in the absence of scale diseconomies, is distributed normally with mean 6.5% and standard deviation 6%. Given the within-objective standard deviations of returns of 7.4 over the past 12 months, in the absence of scale diseconomies, this distribution of alpha would imply (within-objective) persistence coefficients of 0.42.

scale. On the one hand, the confidence intervals around the estimates of scale diseconomies implied by these plausibly exogenous flows are wider than those implied by cross-sectional comparisons of large and small funds. This raises the possibility that diseconomies of scale are larger than previously estimated, perhaps because more-skilled managers are more likely to manage larger funds. On the other hand, the upper bounds of the confidence intervals are too low to significantly change our views about the extent of performance persistence.

Our results suggest that scale diseconomies are not large enough for Berk-Green effects to overturn the conventional interpretation of the mutual fund literature's stylized facts, but this need not diminish interest in the model. The point that successful agents are often given more resources to manage, and these extra resources may overwhelm their ability to perform, is clearly more general than mutual funds. The "Peter Principle" (Peter and Hull, 1969), which argues that managers are promoted to the point where they are no longer effective, is one popular expression of this general idea.

Indeed, there are several reasons that mutual funds may be a setting in which these effects are particularly modest. First, mutual funds trade in public financial markets that are extremely competitive and efficient, relative to other financial markets (such as private equity) and relative to the product markets in which most firms operate. This fact should mitigate the return to all but the highest levels of skill. It should also minimize the cost associated with using incremental flows to deepen existing positions, which is what we find. Second, mutual funds are constrained in their ability to pay for performance, and often lose current and potential managerial talent to less regulated vehicles. Third, mutual fund returns arguably have a larger luck component than performance measures in other fields, complicating the inference of skill from performance. Fourth, to the extent that mutual funds are primarily marketed to unsophisticated investors, the reallocation of resources to more skilled managers will be less efficient than in other settings. And finally, there are fixed costs in asset management, which may explain why we find little evidence of net diseconomies of scale. Applying the insights of the Berk-Green model to settings where skill translates more readily into performance, where true performance (and not luck) is more readily matched with extra resources, and where extra resources have greater negative impact on performance is likely to be a fruitful direction for further work.

REFERENCES

Berk, Jonathan, and Richard Green, 2004, “Mutual Fund Flows and Performance in Rational Markets,” *Journal of Political Economy* 112, 1269–1295.

Berk, Jonathan, 2005, “Five Myths of Active Portfolio Management,” UC-Berkeley mimeo. http://faculty.haas.berkeley.edu/berk/papers/myth.pdf.

Bessembinder, Hendrik and William Maxwell, 2008, “Transparency and the Corporate Bond Market,” *Journal of Economic Perspectives* 22, 217-234.

Blake, Christopher and Matthew Morey, 2003, “Morningstar Ratings and Mutual Fund Performance,” *Journal of Financial and Quantitative Analysis* 35, 451-483.

Blume, Marshall, 1998, “An Anatomy of Morningstar Ratings,” *Financial Analysts Journal* 54, 19-27.

Card, David, 1999, “The Causal Effect of Education on Earnings,” *Handbook of Labor Economics*, Vol 3, Part 1, ed. Orley Ashenfelter and David Card, North Holland.

Carhart, Mark, 1997, “On Persistence in Mutual Fund Performance,” *Journal of Finance* 52, 57–82.

Chevalier, Judith, and Glenn Ellison, 1997, “Risk Taking by Mutual Funds as a Response to Incentives,” *Journal of Political Economy* 105, 1167–1200.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey Kubik, 2004, “Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization,” *American Economic Review* 94, 1276–1302.

Chen, Joseph, Harrison Hong, and Jeffrey Kubik, 2008, “Outsourcing Mutual Fund Management: Firm Boundaries, Incentives and Performance,” Working Paper.

Cici, Gjergi, Scott Gibson, John Merrick, 2011, "Missing the Marks? Dispersion in Corporate Bond Valuations Across Mutual Funds," *Journal of Financial Economics* 101, 206-226.

Del Guercio, Diane, and Paula Tkac, 2008, “Star Power: The Effect of Morningstar Ratings on Mutual Fund Flow,” *Journal of Financial and Quantitative*

Analysis 43, 907–936.

Doll, Richard, 1998, “Uncovering the Effects of Smoking: Historical Perspective,” *Statistical Methods in Medical Research* 7, 87-117.

Edelen, Roger, 1999, “Investor Flows and the Assessed Performance of Open-end Fund Managers,” *Journal of Financial Economics* 53, 439-466.

Edelen, Roger, Richard Evans, and Gregory Kadlec, 2007, “Scale Effects in Mutual Fund Performance: The Role of Trading Costs,” Working Paper.

Evans, Richard, 2010, “Mutual Fund Incubation,” *Journal of Finance* 65, 1581-1611.

Gutierrez, Ro, William Maxwell, and Danielle Xu, 2009, "On Diseconomies of Scale and Persistent Performance in Corporate-Bond Mutual Funds", Working Paper.

Iliev, Peter, 2010, “The Effect of SOX Section 404: Costs, Earnings Quality, and Stock Prices,” *Journal of Finance* 65, 1163–1196.

Imbens, Guido, and Thomas Lemieux, 2008, “Regression Discontinuity Designs: A Guide to Practice,” *Journal of Econometrics* 142, 615-635.

Jensen, Michael, 1968, “The Performance of Mutual Funds in the Period 1945–1964,” *Journal of Finance* 23, 389–416.

Lee, David, and Thomas Lemieux, 2009, “Regression Discontinuity Designs in Economics,” NBER Working Paper #14723.

Massimo, Massa, Jonathan Reuter, and Eric Zitzewitz, 2010, “When Should Firms Share Credit with Employees? Evidence from Anonymously Managed Mutual Funds ,” *Journal of Financial Economics* 95, 400–424.

Peter, Laurence and Raymond Hull, 1969, *The Peter Principle: Why Things Always Go Wrong*, New York: William Morrow and Company.

Pollet, Joshua, and Mungo Wilson, 2008, “How Does Size Affect Mutual Fund Behavior?” *Journal of Finance* 63, 2941-2969.

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

Reuter, Jonathan, and Eric Zitzewitz, 2006, “Do Ads Influence Editors? Advertising and Bias in the Financial Media,” *Quarterly Journal of Economics* 121, 197-227.

Sirri, Erik, and Peter Tufano, 1998, “Costly Search and Mutual Fund Flows,” *Journal of Finance* 53, 1589–1622.

Yan, Xuemin, 2008, "Liquidity, Investment Style, and the Relation between Fund Size and Fund Performance," *Journal of Financial and Quantitative Analysis* 43, 741-768.

Zitzewitz, Eric, 2011, “Paired Bond Trades,” working paper.

Appendix

This appendix has two subsections. First, we present tests for manipulation of the forcing variable. Second, we examine the robustness of our results to alternative bandwidths around the rating thresholds.

A.1. Tests for manipulation

A common concern in regression discontinuity design studies is that agents may be able to manipulate the forcing variable. In our context, the concern would be that funds with more skilled managers might be better at managing their returns (e.g., by time varying the riskiness of their portfolios) in order to maximize their chances of placing just above a Morningstar boundary. If manager skill is correlated with the probability of such manipulation succeeding, this could create a discontinuity in skill at the boundary, confounding our use of the boundary to identify scale diseconomies.

Our prior is that this type of manipulation would be very difficult to do successfully for several reasons. First, the cutoffs for additional stars are not fixed, but depend on the performance of the other funds in a Morningstar category. The median fund*month is in a category with 131 other funds, and so tracking and forecasting the performance of the others funds that are close to a cutoff would be non-trivial. Second, attempting to manage returns through the variation of risk (e.g., as in Chevalier and Ellison, 1997) would be difficult for most mutual funds, which often are restricted in their use of leverage or short selling. Manipulating returns through discretion in the valuation of less-illiquid assets like corporate or municipal bonds might be more feasible (see, e.g., Cici, Gibson, and Merrick, 2011), and so we run separate manipulation tests for these asset categories. Third, the Morningstar ranking methodology is complicated. We are therefore more concerned about manipulation before June 2002, when the methodology became substantially more complicated, and conduct separate tests for the pre-2002 period.

Given these concerns, we are therefore not that surprised that our tests fail to find evidence of successful manipulation, even in the time periods and asset classes where our concerns are the greatest. We conduct three types of tests: 1) McCrary (2008) tests for discontinuities in the density of the forcing variables at rating cutoffs, 2) tests for discontinuities in the control variables at the rating cutoffs, and 3) fund-level tests for persistence in the discontinuity variable (controlling for the forcing variable). The first two tests are fairly standard in regression discontinuity studies, while the third addresses a form of manipulation that would be particularly problematic for our methodology.

A.1.1 McCrary tests for discontinuities

McCrary outlines a simple test for manipulation of the forcing variable. If agents successfully manipulate the forcing variable to just exceed a cutoff, then we should observe a discontinuity in the density of the variable at the cutoff. As successful manipulators would be more common just above the cutoff, to the extent that skill in manipulation was correlated with skill in fund management, this could create a discontinuity in skill at the cutoffs, and thus be a threat to identification.

The McCrary test involves defining a bandwidth, calculating the number of observations in each bin, and testing whether the number of observations in each bin changes discretely at the boundary. Morningstar ratings are defined based on within-category percentile rankings on risk-adjusted returns. Because the density of a percentile variable cannot vary, we test for discontinuities in the density of Morningstar's measure of risk-adjusted returns. For simplicity, we test for discontinuities at the boundaries of the 3, 5 and 10-year star ratings, without considering whether a boundary is pivotal for a fund's overall rating. As in the main analysis, we restrict attention to a fund's largest share class, which for the reasons discussed above is the rating funds are most likely to have an interest in manipulating.

In order to aggregate results over investment categories with different standard deviations of returns, we first standardize returns into z-scores within each objective*month. If returns were normally distributed, the 4 cutoffs would be at z-scores of approximately -1.28, -0.45, 0.45, and 1.28. In practice, the average and standard deviation z-score boundaries are -1.13, -0.29, 0.46, and 1.22, reflecting some excess kurtosis and negative skew. The standard deviations are 0.22, 0.15, 0.15, and 0.24, reflecting some variation around these averages across objective*months.

After identifying the funds that are just above and below each boundary, we place funds into bins that are 0.01 wide in z-score space and begin just after the first fund on each side of the boundary. Figure A1 plots the share of funds in the 10 bins on either side of the four star boundaries for the 3-year ratings (the time horizon which have the largest sample size). The standard normal density is shown for comparison. There is no evidence of a discontinuity in density, and this is confirmed by statistically insignificant coefficients on a discontinuity variable in a regression with a local linear control. Results are similar for the 5-year and 10-year ratings. They are also similar when we split the sample based on pre/post-2002 and bond/stock funds.

A.1.2. Tests for discontinuities in control and lagged dependent variables

A second commonly used test for manipulation of the forcing variable is to test whether control and lagged dependent variables exhibit discontinuities. Table A1 replicates the regressions in Table 3 with alternative dependent variables. While most of the control variables covary with the forcing variable, there is no evidence discontinuities at the rating boundaries. The fact that there are no discontinuities in these pre-determined observable variables increases our confidence that there is also no discontinuity in manager skill, the key unobservable variable our approach is attempting to control for.

A.1.3. Tests for persistence in the discontinuity variable

As a final test for successful manipulation of Morningstar ratings, we ask whether certain funds are able to persistently place just above rating thresholds. Morningstar ratings are based on risk-adjusted returns over multi-period windows (36, 60, and 120 months), and so the month-to-month change in a fund's percentile ranking is fairly small (with standard deviations of 6.6, 5.5, and 4.7 percentile points, respectively). As a result, these tests require controlling for the forcing variable non-linearly.

To see this, suppose that if a fund's percentile ranking change is drawn randomly each month from a p.d.f. f(x), with a peak at x=0. The probability that a fund will exceed a threshold next month will be given by F(z), where z is the current-month distance above the threshold and F is the c.d.f. Given the peak in f() at 0, this probability F(z) will be non-linear in z with a slope that peaks at zero. Especially if the spread in f() is small, controlling for the current-month forcing variable linearly when predicting the future discontinuity could lead one to falsely conclude that there is persistence in the discontinuity variable.

As a result, we use a logit model to predict the future discontinuity variable as a function of the current-month discontinuity and forcing variable, as well as the control variables used in Table 3. Table A2 reports the coefficients on the current-month discontinuity and forcing variable for models predicting the discontinuity variable at the 1, 3, and 6 month horizon. In general, the coefficients on the current-month discontinuity variable are statistically and economically insignificant. In particular, we can usually rule out that being above the boundary this month raises the future probability by more than a 2-percentile increment in the forcing variable.

A.2 Robustness to bandwidth

It is common in regression discontinuity studies to check the robustness of the results to alternative choices about the bandwidth analyzed around the discontinuities. The results presented above use a bandwidth of 5 percentiles in either direction of the cutoff for a Morningstar rating. Table A3 replicates selected results from Tables 4 using alternative bandwidths.

The results do not appear sensitive to the choice of bandwidth. Because narrower bandwidths reduce the number of observations included, standard errors rise, particularly if the bandwidth is narrowed to 2 percentiles. Likewise, standard errors decline slightly as the bandwidth widens. The main conclusions are robust to bandwidth: 1) star boundaries are accompanied by discontinuities in flows, 2) discontinuities in flows are larger for the 4/5-star boundary, and 3) discontinuities in flows are rarely accompanied by discontinuities in future returns.

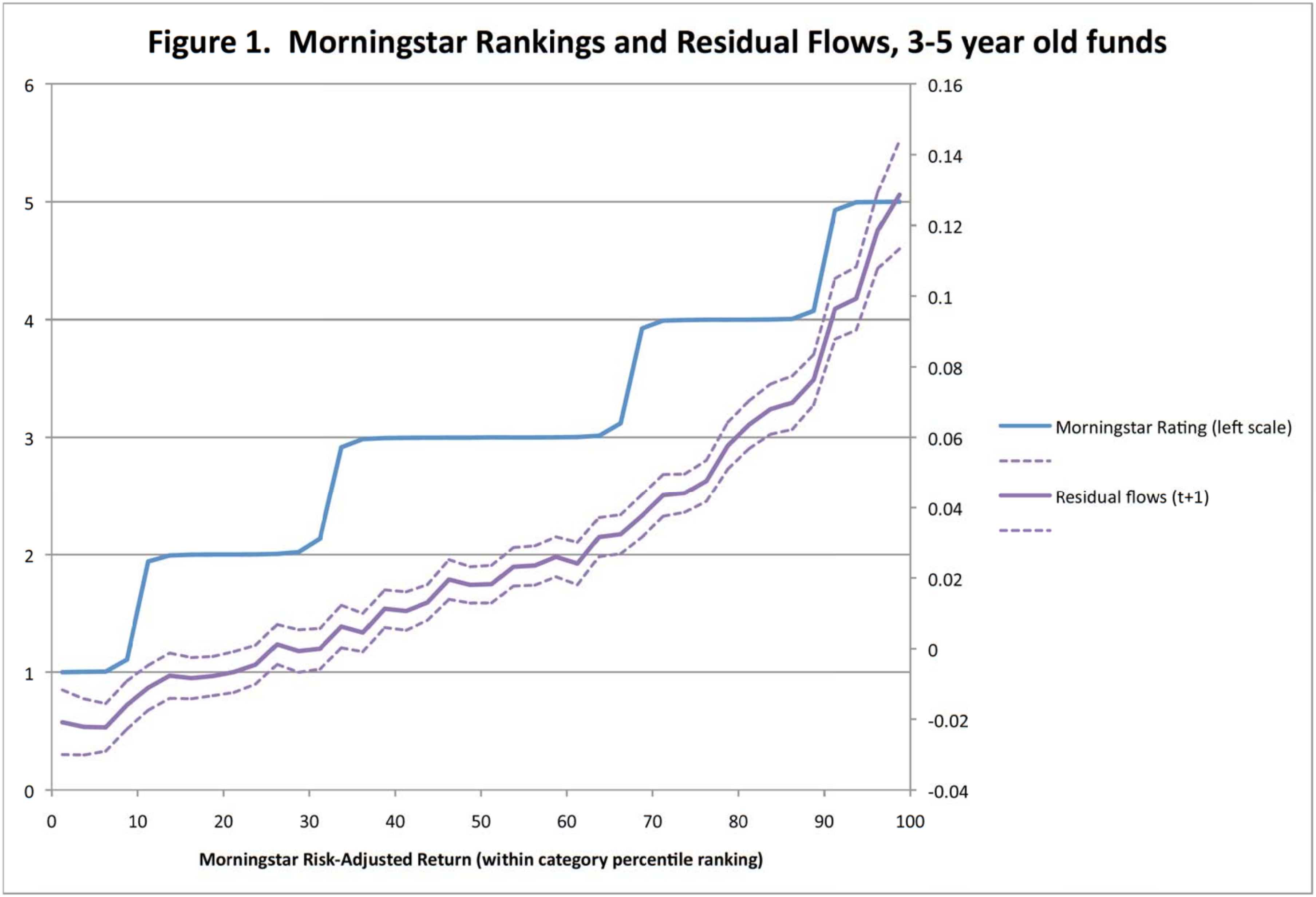


Figure 1. Morningstar Rankings and Residual Flows, 3-5 year old funds
6
5
4
3
2
1
0
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
0
10
20
30
40
50
60
70
80
90
100
Morningstar Risk-Adjusted Return (within category percentile ranking)
Morningstar Rating (left scale)
Residual flows (t+1)

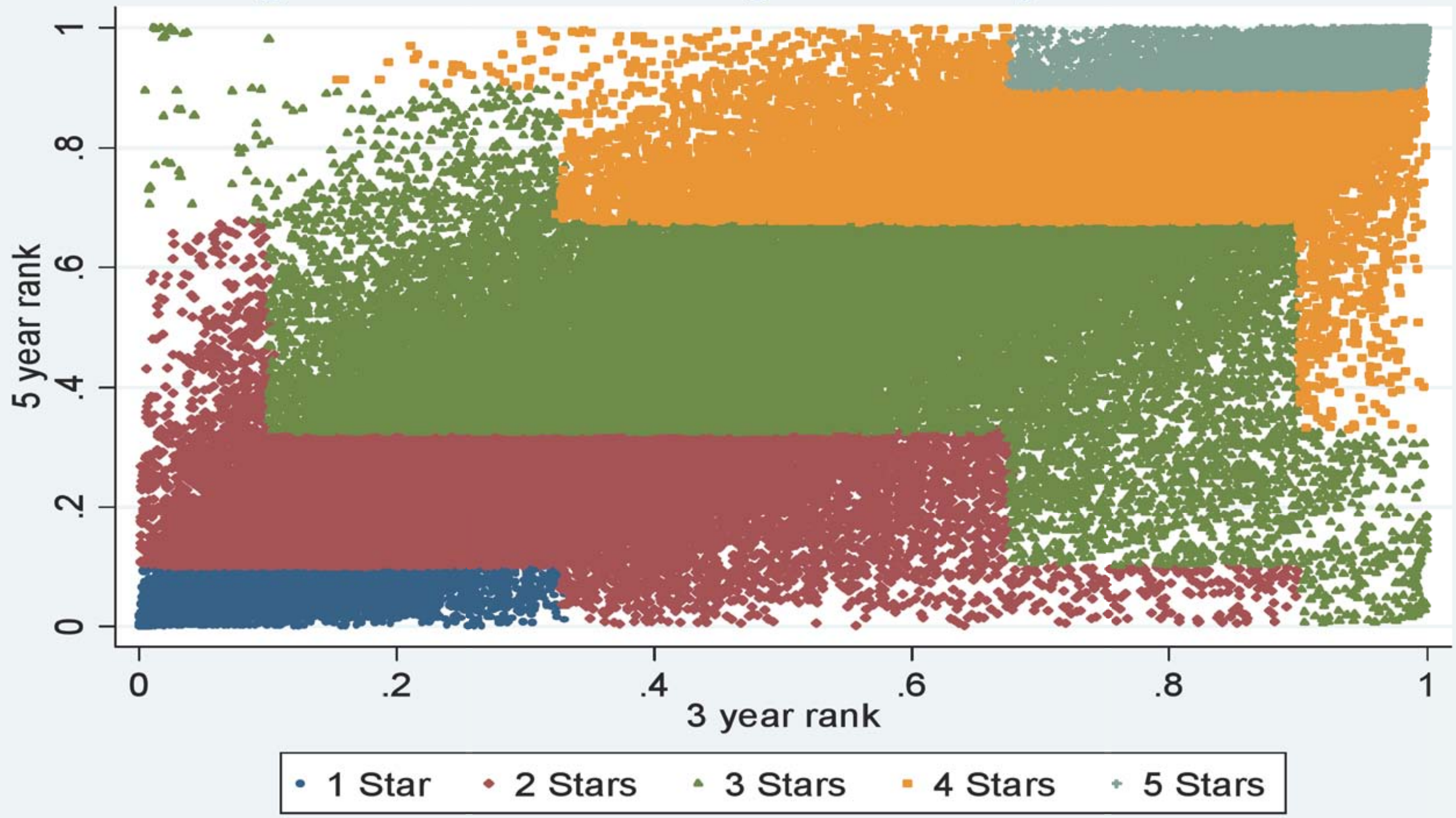


Figure 2. Overall ratings for 5-10 year-old funds

Figure 3A. Next-month residual flows for funds around Morningstar rating boundaries

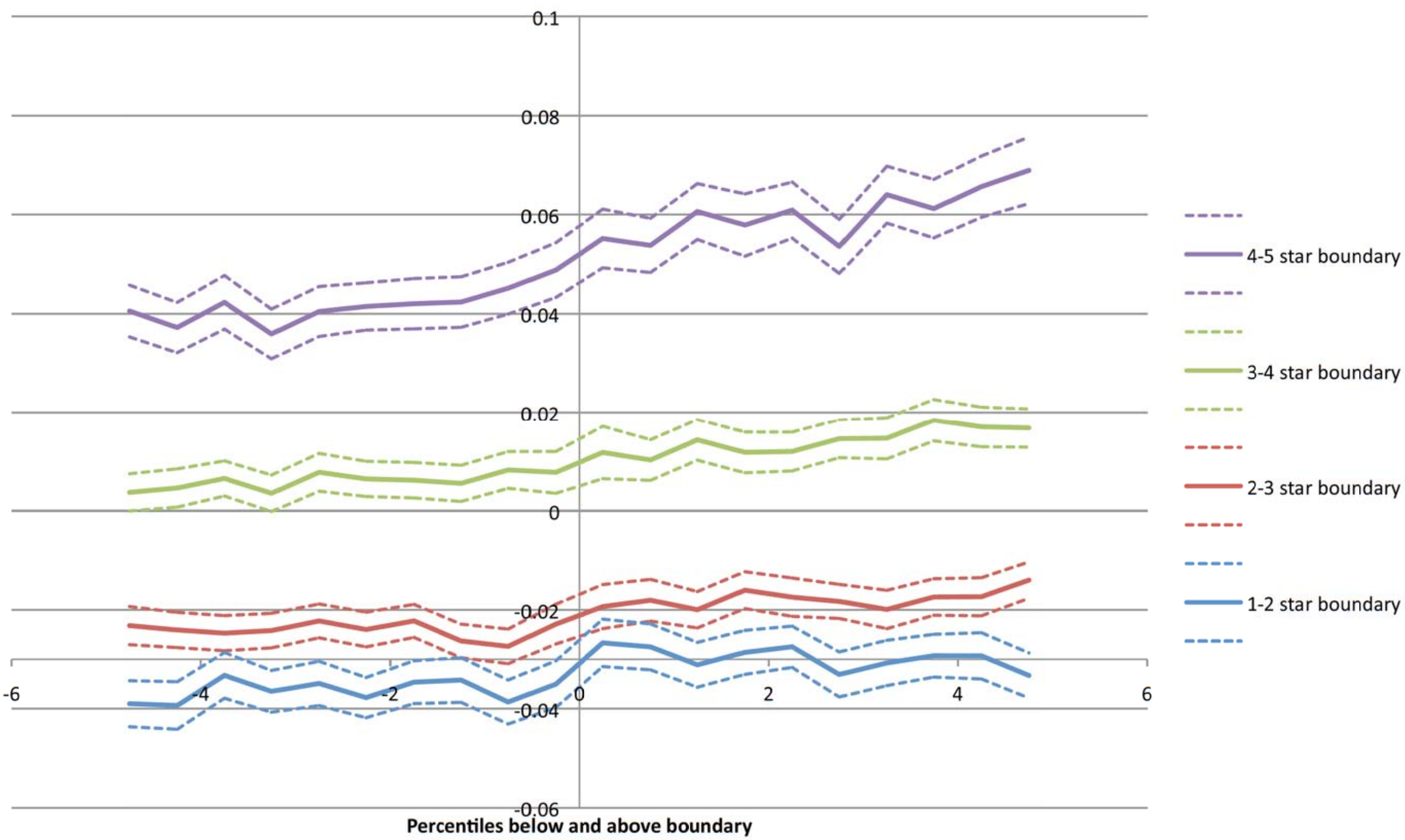
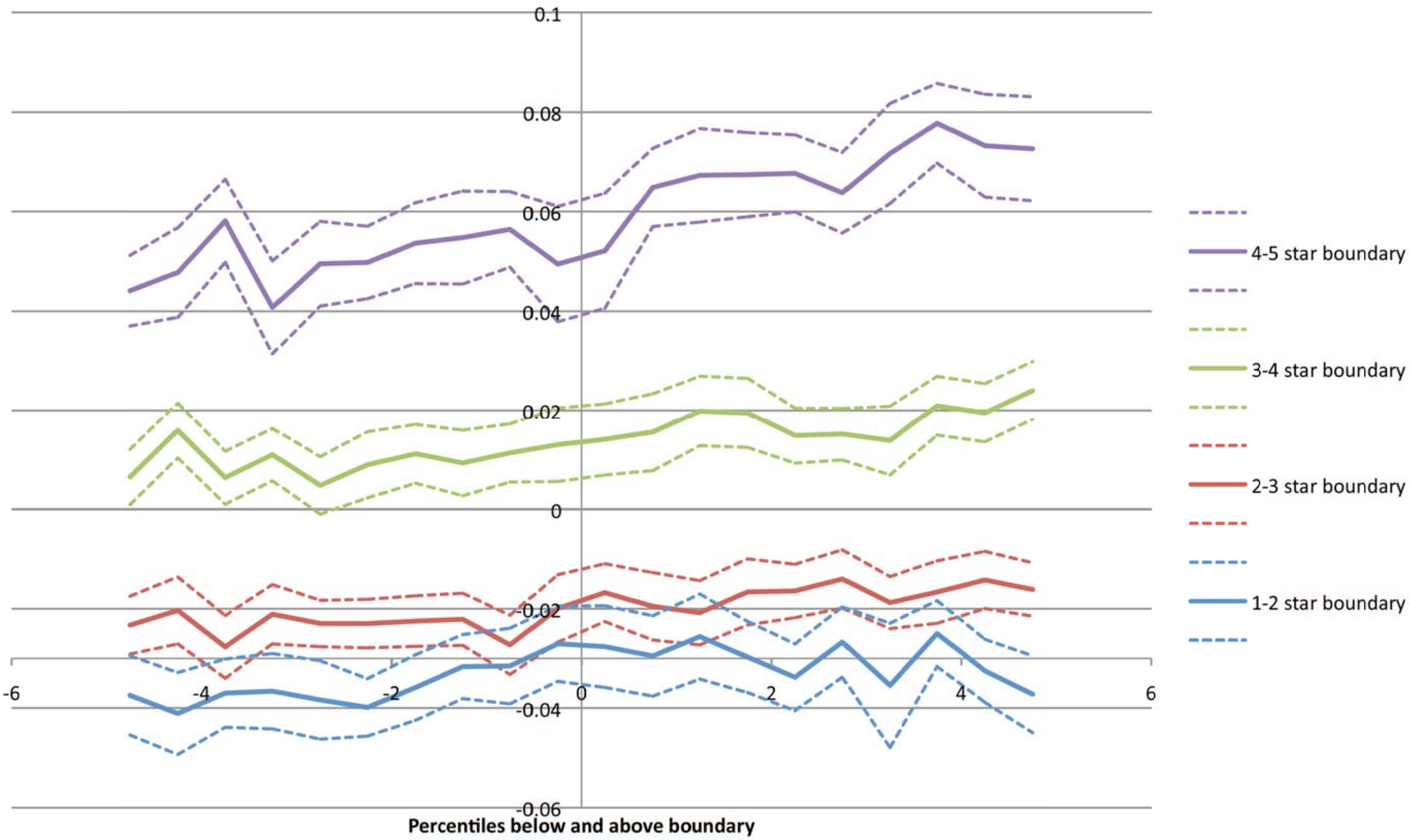


Figure 3B. Falsification: Current-month residual flows for funds around Morningstar rating boundaries



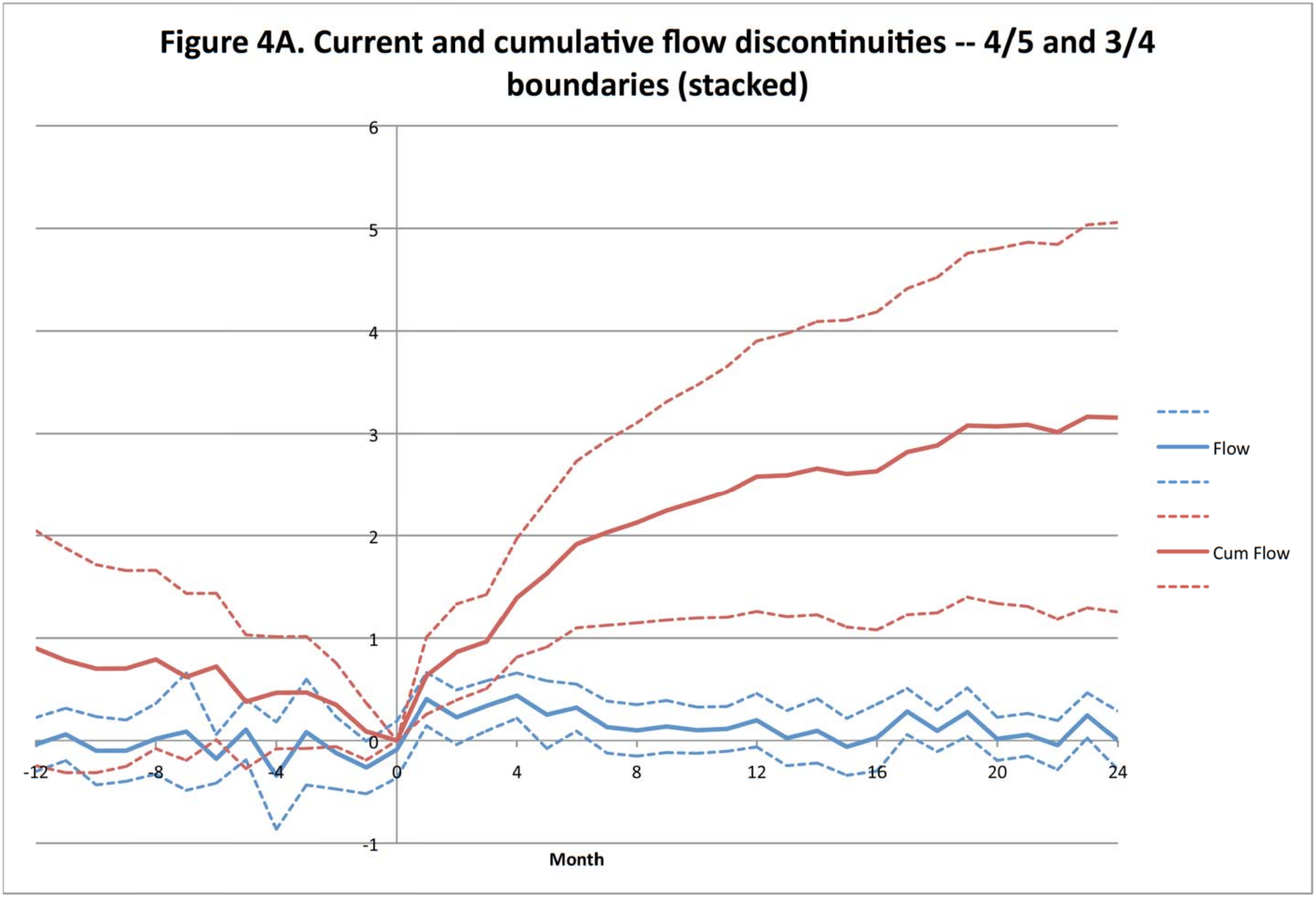


Figure 4A. Current and cumulative flow discontinuities -- 4/5 and 3/4 boundaries (stacked)
6
5
4
3
2
1
0
-1
-12
-8
-4
0
4
8
12
16
20
24
Month
Flow
Cum Flow

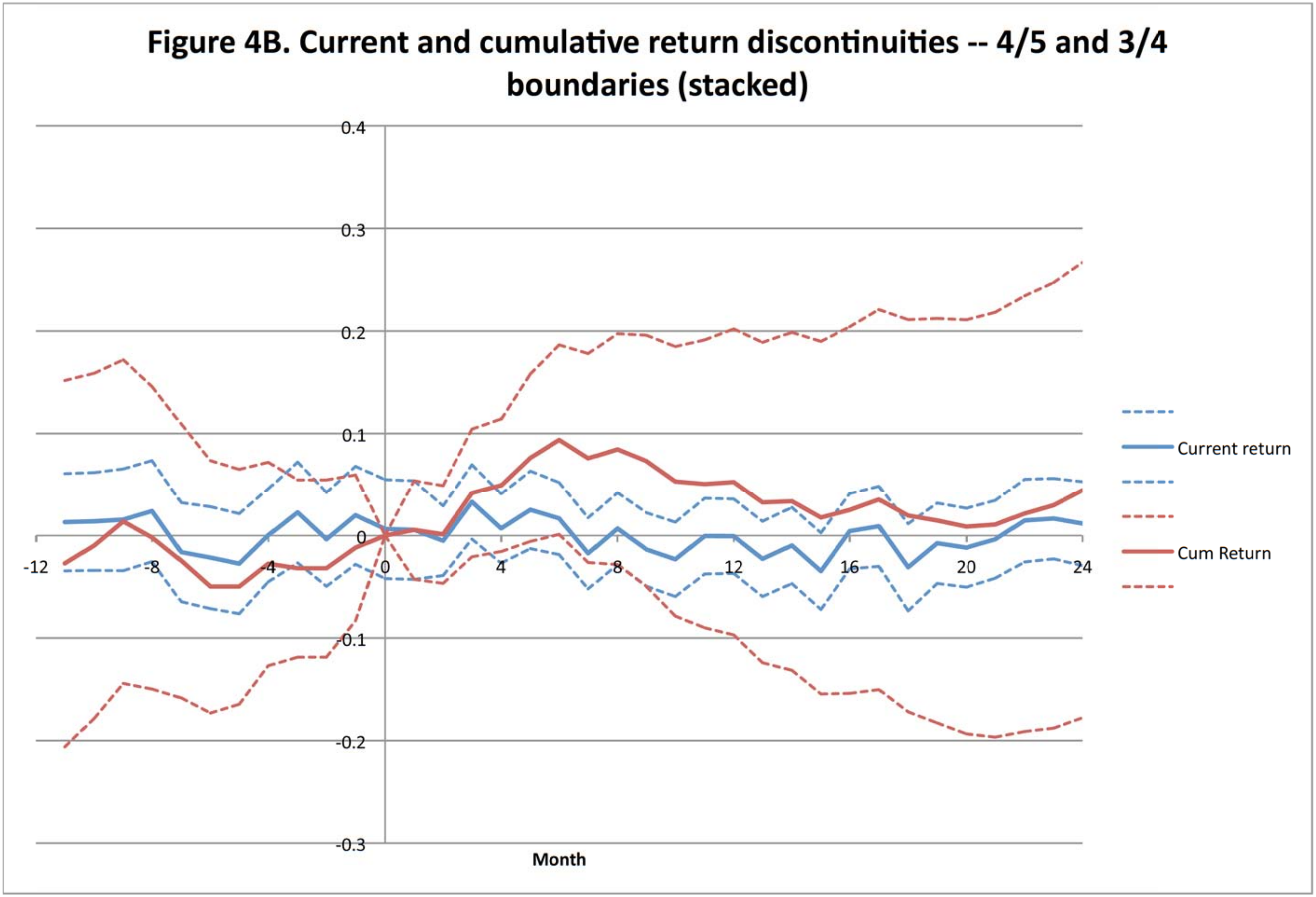


Figure 4B. Current and cumulative return discontinuities -- 4/5 and 3/4 boundaries (stacked)
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-12
-8
-4
0
4
8
12
16
20
24
Month
Current return
Cum Return

Table 1. Summary statistics

Summary statistics are reported for portfolio*month combinations in which at least one share class receives a Morningstar rating, which requires at least three years of history. Fund characteristics are the average characteristics of the funds' share classes, weighted by prior-month assets.

	All funds		By (asset-weighted average) Morningstar rating				
	Mean	SD	1 star	2 stars	3 stars	4 stars	5 stars
Number of portfolio*months	491,863		35,419	101,102	178,293	125,422	51,627
Returns (cumulative log percentage points, adjusted for category mean)							
Percent surviving to t+24 months?	86	35	73	80	86	91	93
Return (t)	0.0	1.8	-0.6	-0.2	0.0	0.2	0.3
Return (t+1)	0.0	1.8	-0.2	0.0	0.0	0.0	0.1
Return (t+1 to t+6)	0.0	5.1	-0.9	-0.2	0.0	0.2	0.4
Return (t+1 to t+12)	0.0	7.4	-1.4	-0.2	0.1	0.3	0.3
Return (t+1 to t+24)	0.0	10.0	-2.1	-0.2	0.1	0.4	0.4
Flows (cumulative log percentage points, adjusted for category mean)							
Flow (t)	0.0	10.6	-1.6	-1.0	-0.2	0.7	2.2
Flow (t+1)	0.0	10.3	-1.4	-0.9	-0.2	0.6	1.9
Flow (t+1 to t+6)	0.0	25.5	-7.1	-4.6	-1.2	3.3	9.8
Flow (t+ 1 to t+12)	0.0	37.8	-12.0	-8.1	-2.0	5.9	16.1
Flow (t+1 to t+24)	0.0	55.2	-17.7	-11.8	-2.9	8.9	23.0
Other fund characteristics							
Ln(Portfolio TNA)	5.2	1.8	4.2	4.8	5.2	5.6	5.9
Ln(Family TNA)	8.9	2.3	8.0	8.8	8.9	9.2	9.2
Expense ratio	1.20	0.72	1.74	1.37	1.15	1.03	1.03
Expense ratio (t+12)	1.19	0.71	1.77	1.38	1.16	1.03	1.02
Percent with any load	69	46	77	77	72	61	55
Portfolio turnover (%)	100	157	150	108	92	87	104
Number of share classes	2.4	1.6	2.3	2.5	2.5	2.3	2.2
Percent of assets in largest share class	84.5	19.4	82.6	82.1	84.5	86.2	86.8
Morningstar ratings							
Percent with same rating for all share classes	73	44	82	70	70	73	83
Average rating (t)	3.10	1.06	1.03	2.01	2.99	3.97	4.97
Average rating (t + 36 months)	3.07	1.02	2.16	2.62	2.99	3.40	3.63
Average risk-adjusted return percentile score							
3 year	51	29	8	25	48	72	91
5 year	52	29	5	22	49	76	93
10 year	51	28	6	23	48	74	91
Percent of portfolios with:							
5 year rating (5-9 year-old funds)	83	37	78	84	85	84	78
10 year rating (10+ year-old funds)	46	50	39	47	49	46	40

Table 2. Estimated future inflow discontinuity at Morningstar rating boundaries -- share class level
Dependent variable: log net flows (in percent)

The unit of observation is share class i in month t. The sample is restricted to share classes of actively managed funds that are at least three years old, because younger share classes are not eligible for a Morningstar rating. The sample for each regression is further restricted to share classes with within-category performance rankings within 5 percentiles on either side of a ranking boundary. The stacked models include all 4 ranking boundaries; the other models include only the indicated boundary. All regressions include a local linear control for within-category percentile ranking and and a discontinuity variable that equals one when the share classes' within-category percentile ranking is above the boundary. Baseline regressions also include controls for lagged log size of the share class, portfolio, and family, expense ratio, portfolio turnover, and loads (front, deferred, and trailing), and category*month fixed effects. "Additional controls" include controls for Morningstar's measure of risk-adjusted returns, lagged log returns from t-12 to t-1, t-24 to t-13, and t-36 to t-25, and lagged log inflows from t-12 to t-1 and t-3 to t-1. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

		Next-month flows				Current-month flows			
Boundary		Baseline		Additional Controls		Baseline		Additional Controls	
Stacked	Discontinuity	0.518	***	0.432	***	-0.006		0.033	
		(0.062)		(0.062)		(0.091)		(0.090)	
	Local linear control	0.007		0.003		0.054	***	0.037	**
		(0.011)		(0.010)		(0.016)		(0.016)	
4/5 stars	Discontinuity	0.946	***	0.880	***	0.299		0.367	
		(0.141)		(0.143)		(0.258)		(0.268)	
	Local linear control	0.046	*	0.022		0.070		0.020	
		(0.025)		(0.026)		(0.046)		(0.048)	
3/4 stars	Discontinuity	0.854	***	0.671	***	-0.196		-0.151	
		(0.130)		(0.132)		(0.165)		(0.161)	
	Local linear control	-0.040	*	-0.036	*	0.072	**	0.053	*
		(0.022)		(0.022)		(0.028)		(0.028)	
2/3 stars	Discontinuity	0.610	***	0.514	***	0.232		0.247	*
		(0.108)		(0.106)		(0.159)		(0.154)	
	Local linear control	-0.027		-0.037	*	0.011		-0.001	
		(0.019)		(0.019)		(0.028)		(0.027)	
1/2 stars	Discontinuity	0.337	**	0.312	**	-0.225		-0.180	
		(0.137)		(0.138)		(0.197)		(0.186)	
	Local linear control	0.009		0.001		0.076	**	0.057	*
		(0.026)		(0.026)		(0.033)		(0.031)	

Table 3. Estimated future inflow discontinuity at Morningstar rating boundaries -- portfolio level
Dependent variable: log net flows (in percent)

This table repeats the analysis in Table 2, but the unit of observation is actively managed fund j in month t. The local linear control is the within-category percentile ranking for the largest share class, and the discontinuity measure indicates whether the portfolio's largest share class is on the positive side of the rating boundary. Other variables are aggregated to the portfolio level, weighted by prior-month assets. The sample is restricted to portfolios for which the largest share classes' percentile ranking is within 5 percentiles of the rating boundary. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

		Next-month flows				Falsification: current-month flows			
Boundary		Baseline		Additional Controls		Baseline		Additional Controls	
Stacked	Discontinuity	0.586	***	0.406	***	-0.078		-0.088	
		(0.132)		(0.134)		(0.109)		(0.110)	
	Local linear control	0.063	**	0.048	*	0.088	***	0.070	***
		(0.024)		(0.025)		(0.019)		(0.020)	
4/5 stars	Discontinuity	0.587	**	0.520	*	-0.040		-0.009	
		(0.268)		(0.282)		(0.203)		(0.213)	
	Local linear control	0.213	***	0.152	***	0.147	***	0.112	***
		(0.054)		(0.055)		(0.039)		(0.040)	
3/4 stars	Discontinuity	0.852	***	0.613	**	-0.239		-0.223	
		(0.264)		(0.269)		(0.192)		(0.193)	
	Local linear control	0.008		-0.002		0.099	***	0.075	**
		(0.043)		(0.044)		(0.030)		(0.030)	
2/3 stars	Discontinuity	0.513	**	0.385		0.045		0.043	
		(0.250)		(0.262)		(0.208)		(0.218)	
	Local linear control	0.004		0.008		0.002		-0.002	
		(0.043)		(0.045)		(0.036)		(0.037)	
1/2 stars	Discontinuity	0.099		-0.217		-0.054		-0.072	
		(0.372)		(0.386)		(0.328)		(0.335)	
	Local linear control	0.061		0.006		0.109	*	0.091	
		(0.079)		(0.085)		(0.062)		(0.066)	

Table 4. Discontinuities in cumulative flows and returns -- portfolio level
Dependent variable: log net flows or log net returns (in percentage points)

The table repeats the regressions in Table 3 for different time horizons. The dependent variable is the cumulative fund-level log flows or fund-level log returns, calculated either from the ranking month to a future month, or from a past month to the ranking month. As in Table 3, all regressions include the discontinuity and within-category percentile ranking variables for the largest share class. They also include the full set of fund-level controls, including the "additional controls" for past returns and inflows, and fixed effects for category*month combinations. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

	Flows				Returns			
Month	All boundaries (stacked)	4/5 boundary	3/4 boundary	4/5 and 3/4 (stacked)	All boundaries (stacked)	4/5 boundary	3/4 boundary	4/5 and 3/4 (stacked)
-12	-0.43	-0.78	-0.09	-0.45	0.01	-0.04	-0.02	-0.01
	(0.44)	(0.92)	(0.77)	(0.59)	(0.02)	(0.04)	(0.03)	(0.03)
-9	-0.26	-0.62	-0.12	-0.38	0.00	-0.02	-0.12 **	-0.03
	(0.37)	(0.77)	(0.65)	(0.50)	(0.03)	(0.07)	(0.05)	(0.04)
-6	-0.19	-0.60	-0.07	-0.34	0.02	0.03	0.02	0.01
	(0.29)	(0.67)	(0.47)	(0.40)	(0.04)	(0.07)	(0.05)	(0.04)
-3	-0.11	-0.34	-0.11	-0.25	0.05 *	0.04	0.02	0.01
	(0.20)	(0.44)	(0.31)	(0.27)	(0.03)	(0.06)	(0.05)	(0.04)
0	-0.09	-0.01	-0.22	-0.10	0.01	-0.01	0.03	0.00
	(0.11)	(0.21)	(0.19)	(0.14)	(0.02)	(0.04)	(0.03)	(0.02)
1	0.41 ***	0.52 *	0.61 **	0.63 ***	0.01	0.01	0.00	0.00
	(0.13)	(0.28)	(0.27)	(0.19)	(0.02)	(0.04)	(0.03)	(0.02)
2	0.58 ***	0.70 *	0.88 ***	0.85 ***	0.03	0.05	0.03	0.04
	(0.17)	(0.38)	(0.32)	(0.23)	(0.02)	(0.05)	(0.04)	(0.03)
3	0.70 ***	1.04 ***	0.81 **	0.96 ***	0.01	0.00	0.06	0.05
	(0.17)	(0.38)	(0.32)	(0.23)	(0.03)	(0.05)	(0.04)	(0.03)
6	1.12 ***	2.04 ***	1.74 ***	1.90 ***	0.01	0.06	0.06	0.08
	(0.28)	(0.57)	(0.61)	(0.41)	(0.04)	(0.09)	(0.06)	(0.05)
9	1.42 ***	2.03 ***	2.24 ***	2.24 ***	0.00	-0.03	0.12	0.05
	(0.36)	(0.75)	(0.79)	(0.53)	(0.05)	(0.12)	(0.08)	(0.07)
12	1.56 ***	2.33 **	2.50 **	2.58 ***	-0.04	-0.09	0.11	0.03
	(0.44)	(0.93)	(0.98)	(0.66)	(0.06)	(0.14)	(0.10)	(0.08)
15	1.61 ***	2.27 **	2.65 **	2.61 ***	-0.04	-0.12	0.11	0.03
	(0.50)	(1.04)	(1.12)	(0.75)	(0.07)	(0.15)	(0.11)	(0.09)
18	1.85 ***	2.48 **	2.94 **	2.89 ***	-0.06	-0.09	0.09	0.02
	(0.54)	(1.16)	(1.20)	(0.82)	(0.08)	(0.17)	(0.12)	(0.10)
21	2.09 ***	2.30 *	3.32 ***	3.09 ***	-0.09	-0.06	0.07	0.02
	(0.59)	(1.30)	(1.27)	(0.89)	(0.09)	(0.18)	(0.13)	(0.11)
24	2.30 ***	2.27 *	3.41 **	3.16 ***	-0.08	-0.02	0.08	0.05
	(0.63)	(1.40)	(1.34)	(0.95)	(0.09)	(0.20)	(0.14)	(0.11)

Table 5. Cumulative flow and return effects by asset class -- 4/5 and 3/4 boundaries (stacked)
Dependent variable: log net flows or log net returns (in percentage points)

This table repeats the first-stage and reduced-form regressions in the "4/5 and 3/4 (stacked)" regressions in Table 4 for different time horizons and subsamples of actively managed mutual funds. "All equity" includes "Large-cap equity", "Mid-cap equity", "Small-cap equity", "Sector funds", and "International equity". "All funds" includes "All equity", "Taxable bonds", "Municipal bonds", and a relatively small number of funds that do not fit into these subsamples, such as balanced funds. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

	All funds	All equity	Large-cap equity	Mid-cap equity	Small-cap equity	Sector funds	International equity	Taxable bonds	Municipal bonds
Flow effects									
T+6 months	1.92 ***	2.65 ***	1.10	1.60	2.03	6.19 **	4.51 *	2.31 **	1.10 **
	(0.41)	(0.42)	(0.77)	(1.20)	(1.71)	(2.58)	(2.56)	(0.91)	(0.47)
T+12 months	2.58 ***	3.57 ***	1.73	2.69	-0.64	8.84 **	5.72	2.59 *	2.07 ***
	(0.66)	(0.60)	(1.23)	(1.90)	(3.38)	(3.65)	(4.21)	(1.38)	(0.68)
T+18 months	2.88 ***	3.80 ***	2.08	2.50	-1.95	11.77 **	5.10	3.81 **	2.29 ***
	(0.82)	(0.76)	(1.75)	(2.58)	(3.80)	(4.62)	(4.85)	(1.71)	(0.77)
T+24 months	3.16 ***	3.73 ***	1.90	2.54	-2.69	13.95 **	5.21	4.29 **	2.91 ***
	(0.95)	(0.88)	(2.08)	(2.99)	(4.44)	(6.35)	(5.26)	(2.07)	(0.92)
Return effects									
T+6 months	0.08	0.06	-0.04	0.55 **	-0.29	0.29	0.23	0.00	0.07 **
	(0.05)	(0.07)	(0.11)	(0.28)	(0.28)	(0.43)	(0.25)	(0.05)	(0.03)
T+12 months	0.03	-0.08	-0.08	0.76 *	-0.67	-0.14	-0.15	-0.07	0.09 **
	(0.08)	(0.11)	(0.18)	(0.42)	(0.44)	(0.55)	(0.37)	(0.08)	(0.04)
T+18 months	0.01	-0.19	-0.17	0.80 *	-1.06 *	-0.02	0.00	-0.06	0.10 **
	(0.10)	(0.14)	(0.23)	(0.50)	(0.61)	(0.67)	(0.45)	(0.10)	(0.05)
T+24 months	0.05	-0.22	-0.21	1.19 **	-1.04	0.05	-0.05	-0.06	0.10 *
	(0.11)	(0.15)	(0.27)	(0.55)	(0.71)	(0.79)	(0.50)	(0.11)	(0.06)

Table 6. Size-performance and flow-performance correlations compared with estimated causal effects of flows

This table compares the partial correlation between portfolio size and future performance with the estimated causal effect of inflows on performance. The partial correlation is estimated as the coefficient on portfolio size in a regression of future returns on the variables listed under fund characteristics in Table 1, a control for past-12-month log returns, and category*month fixed effects. The estimated effects of an extra Morningstar star on future inflows and returns are taken from Table 5. The ratio of these estimated effects provides an instrument variables estimate of the causal effect of inflows on (contemporaneous) performance. Hausman tests generally do not reject the null hypothesis that the "Standard OLS" size-performance correlation and "IV" causal effects are equal. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

Specification:		"First Stage"		"Reduced Form"		"IV"		"Standard OLS"		Hausman test
		Causal effect of star on flows		Causal effect of star on returns		IV estimate of flow effect on returns		Portfolio size coefficient		p-value
Horizon	Asset class	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	
6 months	All funds	1.92	(0.41)	0.08	(0.05)	0.04	(0.03)	-0.0011	(0.0001)	0.13
	All equity	2.65	(0.42)	0.06	(0.07)	0.02	(0.03)	-0.0017	(0.0002)	0.33
	Large-cap equity	1.10	(0.77)	-0.04	(0.11)	-0.03	(0.10)	-0.0019	(0.0003)	0.76
	Mid-cap equity	1.60	(1.20)	0.55	(0.28)	0.34	(0.18)	-0.0017	(0.0005)	0.05
	Small-cap equity	2.03	(1.71)	-0.29	(0.28)	-0.14	(0.14)	-0.0027	(0.0007)	0.30
	Sector funds	6.19	(2.58)	0.29	(0.43)	0.05	(0.07)	-0.0017	(0.0008)	0.47
	International equity	4.51	(2.56)	0.23	(0.25)	0.05	(0.06)	-0.0007	(0.0005)	0.37
	Taxable bonds	2.31	(0.91)	0.00	(0.05)	0.00	(0.02)	-0.0001	(0.0001)	0.97
	Munis	1.10	(0.47)	0.07	(0.03)	0.06	(0.03)	-0.0001	(0.0001)	0.04
12 months	All funds	2.58	(0.66)	0.03	(0.08)	0.01	(0.03)	-0.0021	(0.0003)	0.63
	All equity	3.57	(0.60)	-0.08	(0.11)	-0.02	(0.03)	-0.0030	(0.0004)	0.51
	Large-cap equity	1.73	(1.23)	-0.08	(0.18)	-0.05	(0.10)	-0.0034	(0.0005)	0.68
	Mid-cap equity	2.69	(1.90)	0.76	(0.42)	0.28	(0.16)	-0.0033	(0.0009)	0.07
	Small-cap equity	-0.64	(3.38)	-0.67	(0.44)	Not meaningful		-0.0052	(0.0014)	
	Sector funds	8.84	(3.65)	-0.14	(0.55)	-0.02	(0.06)	-0.0027	(0.0016)	0.83
	International equity	5.72	(4.21)	-0.15	(0.37)	-0.03	(0.06)	-0.0009	(0.0009)	0.69
	Taxable bonds	2.59	(1.38)	-0.07	(0.08)	-0.03	(0.03)	-0.0002	(0.0003)	0.35
	Munis	2.07	(0.68)	0.09	(0.04)	0.04	(0.02)	-0.0001	(0.0002)	0.02
24 months	All funds	3.16	(0.95)	0.05	(0.11)	0.02	(0.04)	-0.0030	(0.0005)	0.61
	All equity	3.73	(0.88)	-0.22	(0.15)	-0.06	(0.04)	-0.0045	(0.0007)	0.18
	Large-cap equity	1.90	(2.08)	-0.21	(0.27)	-0.11	(0.14)	-0.0053	(0.0009)	0.46
	Mid-cap equity	2.54	(2.99)	1.19	(0.55)	0.47	(0.22)	-0.0055	(0.0016)	0.03
	Small-cap equity	-2.69	(4.44)	-1.04	(0.71)	Not meaningful		-0.0077	(0.0023)	
	Sector funds	13.95	(6.35)	0.05	(0.79)	0.00	(0.06)	-0.0029	(0.0029)	0.91
	International equity	5.21	(5.26)	-0.05	(0.50)	-0.01	(0.10)	-0.0010	(0.0015)	0.93
	Taxable bonds	4.29	(2.07)	-0.06	(0.11)	-0.01	(0.03)	-0.0003	(0.0005)	0.58
	Munis	2.91	(0.92)	0.10	(0.06)	0.03	(0.02)	-0.0002	(0.0004)	0.09

Table 7. How are Incremental Flows Invested?

This table reports the estimated coefficients on the discontinuity measure for three different specifications. The sample is smaller than in other tables because it is restricted to non-specialty domestic equity funds for which we observe equity holdings, which are disclosed quarterly. The flow and return specifications match the "4/5 and 3/4 (stacked)" first-stage and reduced-form regressions in Table 4. The holdings specification is analogous to the return specification, except that the dependent variable is the natural logarithm of the number of equity holdings. We report estimates for the full sample of non-specialty domestic equity funds for which we observe equity holdings. We also report estimates for two subsamples, based on how fund i's number of equity holdings compares to the median number of equity holdings for funds with the same investment objective, in the same quarter. Standard errors (in parentheses) allow for clustering within funds.

	Sample with Holdings			# Holdings at or above median			# Holdings below median		
Time horizon	Flow	Return	Holdings	Flow	Return	Holdings	Flow	Return	Holdings
T+3 months	1.22	-0.01	-1.22	0.82	-0.05	-0.72	1.98	0.10	-1.43
	(0.68)	(0.13)	(1.06)	(0.77)	(0.17)	(1.01)	(1.32)	(0.23)	(1.44)
	6,951	6,951	6,951	3,575	3,575	3,575	3,376	3,376	3,376
T+6 months	2.94	0.09	-1.67	2.03	0.01	-0.41	4.41	0.28	-2.70
	(1.00)	(0.21)	(1.05)	(1.35)	(0.29)	(1.02)	(1.77)	(0.33)	(1.61)
	6,654	6,654	6,654	3,424	3,424	3,424	3,230	3,230	3,230
T+12 months	3.06	0.10	-1.06	1.66	-0.36	0.19	5.05	0.57	-2.20
	(1.93)	(0.35)	(1.35)	(3.37)	(0.49)	(1.56)	(2.57)	(0.52)	(1.90)
	6,047	6,047	6,047	3,114	3,114	3,114	2,933	2,933	2,933
T+18 months	3.19	0.16	-0.74	2.06	-0.46	2.27	4.78	0.90	-4.06
	(2.62)	(0.45)	(1.35)	(3.99)	(0.62)	(2.26)	(4.03)	(0.70)	(2.47)
	5,452	5,452	5,452	2,843	2,843	2,843	2,609	2,609	2,609
T+24 months	2.52	0.15	-1.80	0.62	-0.20	0.59	4.54	0.92	-4.85
	(3.22)	(0.57)	(1.88)	(5.01)	(0.76)	(2.44)	(4.92)	(0.91)	(2.87)
	4,869	4,869	4,869	2,558	2,558	2,558	2,311	2,311	2,311

Table 8. Adjusting performance persistence for causal effect of flows

This table reports coefficients from regressions of future returns and inflows on past 12-month returns. Regressions include category*month fixed effects and control for the same control variables as in the first-stage and reduced-form regressions in Tables 2-6. The table then reports return persistence coefficients that are corrected for the causal effect of inflows on performance. The corrections are calculated as the product of the flow coefficient (i.e., the extra inflows that accompany a 1% higher return) and the causal effect of an additional 1% of asset inflow on future performance, as reported in Table 6. We calculate the upper and lower bounds of 95% confidence intervals for the corrected persistence coefficient using the standard errors reported in Table 6. For small-cap equity funds, the corrected coefficients are not meaningful because the first stage (flow) estimate is negative.

		Persistence coefficient		Flow coefficient		Corrected persistence coefficient		
Horizon	Asset class	Coef.	S.E.	Coef.	S.E.	Coef.	Lower CI	Upper CI
12 months	All funds	0.091	(0.010)	0.947	(0.041)	0.079	0.022	0.137
	All equity	0.088	(0.011)	0.894	(0.043)	0.108	0.053	0.164
	Large-cap equity	0.081	(0.025)	0.967	(0.117)	0.126	-0.076	0.329
	Mid-cap equity	0.089	(0.020)	0.909	(0.074)	-0.168	-0.452	0.115
	Small-cap equity	0.122	(0.025)	0.948	(0.084)		Not meaningful	
	Sector funds	0.069	(0.023)	0.693	(0.078)	0.080	-0.006	0.167
	International equity	0.064	(0.022)	0.900	(0.089)	0.088	-0.029	0.205
	Taxable bonds	0.115	(0.023)	1.912	(0.197)	0.170	0.053	0.287
	Munis	0.142	(0.042)	1.415	(0.408)	0.080	0.025	0.135

Figure A1. Ploting the density of actual returns around rating boundaries -- 3 year ratings

We standardize fund returns into z-scores within each objective*month. After identifying the funds that are just above and below each boundary, we place funds into bins that are 0.01 wide in z-score space. We plot the share of funds in the 10 bins on either side of the 1/2, 2/3, 3/4, and 4/5 boundaries for the 3-year ratings. We also plot the share implied by the standard normal distribution.

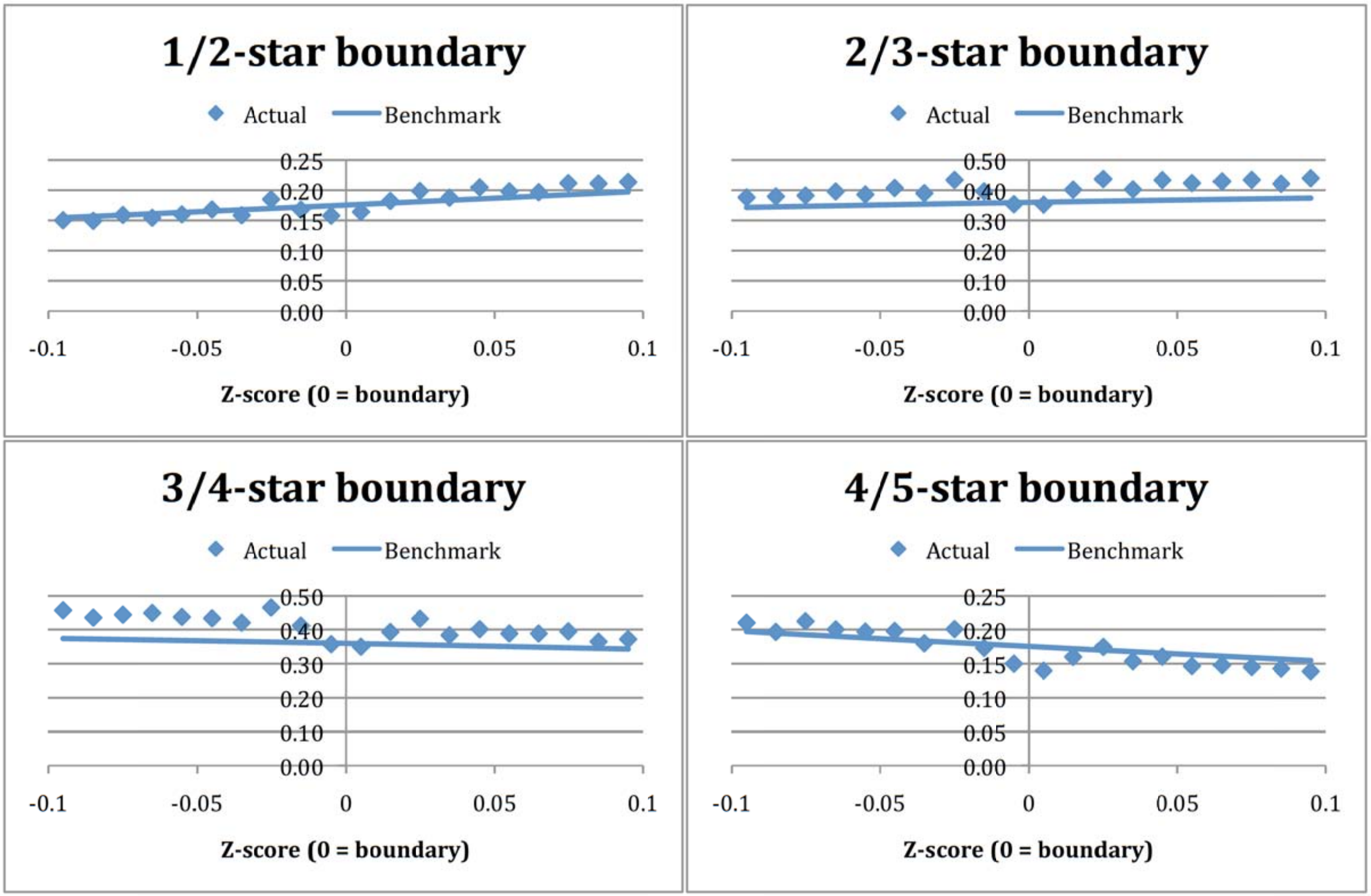


Table A1. Tests for discontinuities in control variables -- portfolio level

This table reports tests for discontinuities at Morningstar ranking borders in the control variables used in the portfolio-level regressions in Tables 3-7. Since these controls are pre-determined at the time of ranking, the Morningstar rating should not have a causal effect on them. The sample is the same as in Table 2. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

	Discontinuity			Local linear control		
Dependent variable	Coef.		S.E.	Coef.		S.E.
Log Portfolio TNA	0.69		(1.58)	1.10	***	(0.31)
Log Family TNA	-3.70	*	(2.16)	1.16	***	(0.42)
Expense ratio	-0.41		(0.38)	0.02		(0.08)
Expense ratio (t+12)	-0.30		(0.41)	0.00		(0.08)
Has load?	-2.29		(5.45)	-0.55	***	(0.10)
Portfolio turnover	-0.12		(0.15)	-0.42	***	(0.03)
Log return (t-12 to t)	-0.15		(0.14)	0.80	***	(0.03)
Morningstar 3-year risk-adjusted return	-0.22		(0.23)	0.73	***	(0.04)

Table A2. Logit regressions predicting future rating

This table reports coefficients from logit regressions predicting future Morningstar ratings using the same specification as in Tables 3-6. For funds at the boundary between s and s+1 stars, the dependent variable is defined as 1 if the fund receives s+1 or more stars in the future. Data from all four boundaries are stacked. Regressions with controls include the full set of controls used in the "additional controls" specification in Table 3. Standard errors (in parentheses) allow for clustering within funds. Statistical significance at the 10-percent, 5-percent, and 1-percent level in two-sided tests is denoted by *, **, and ***.

Time horizon	Controls	Discontinuity	Local linear control
T+1 month	No	0.54	0.425
		(3.41)	(0.007)
	Yes	0.56	0.499
		(5.87)	(0.013)
T+3 month	No	0.25	0.220
		(3.40)	(0.006)
	Yes	0.19	0.283
		(5.81)	(0.012)
T+6 month	No	0.13	0.119
		(3.87)	(0.007)
	Yes	0.13	0.152
		(6.87)	(0.013)

Table A3. Robustness of results to bandwidth

In this table, we report first-stage and reduced-form estimates based on alternative bandwidths. We focus on flows and returns in month t+12. The estimates based on the 0.05 bandwidth match those reported in Table 4. Standard errors (in parentheses) allow for clustering within funds.

		Bandwidth (in percentile space; 0.05 is used in main tables)						
		0.02	0.03	0.04	0.05	0.06	0.07	0.08
Flows, all funds, all boundaries	Discontinuity	1.77	1.78	1.75	1.56	1.27	1.38	1.44
(Table 4, Col 1)		(0.71)	(0.56)	(0.48)	(0.44)	(0.35)	(0.34)	(0.34)
	Running variable	-0.12	-0.05	-0.03	0.08	0.09	0.08	0.07
		(0.27)	(0.15)	(0.10)	(0.08)	(0.07)	(0.07)	(0.07)
	Observations	58,004	88,273	118,225	147,467	161,909	175,755	189,588
Flows, all funds, 4/5 and 3/4 (stacked)	Discontinuity	2.52	2.62	2.71	2.58	2.47	2.21	2.00
(Table 4, Col 4)		(1.12)	(0.89)	(0.74)	(0.66)	(0.52)	(0.53)	(0.51)
	Running variable	0.15	0.10	0.07	0.18	0.22	0.20	0.20
		(0.41)	(0.24)	(0.15)	(0.11)	(0.11)	(0.11)	(0.10)
	Observations	32,030	48,844	65,286	81,310	89,307	96,853	104,422
Returns, all funds, all boundaries	Discontinuity	0.08	0.07	-0.02	-0.03	-0.03	-0.03	-0.02
(Table 4, Col 5)		(0.09)	(0.07)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)
	Running variable	-0.01	-0.02	0.01	0.01	0.01	0.01	0.01
		(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	Observations	58,016	88,294	118,254	147,505	161,951	175,801	189,635
Returns, all funds, 4/5 and 3/4 (stacked)	Discontinuity	0.11	0.12	0.05	0.03	0.02	0.00	0.00
(Table 4, Col 8)		(0.11)	(0.09)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)
	Running variable	-0.02	-0.02	-0.01	0.00	0.00	0.00	0.00
		(0.05)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	Observations	32,032	48,851	65,296	81,324	89,322	96,869	104,439