From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2018

Essays on Portfolio Choice

Branikas, Ioannis

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp019w0325748

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

ESSAYS ON PORTFOLIO CHOICE

IOANNIS BRANIKAS

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF ECONOMICS
ADVISOR: JAKUB KASTL

JUNE 2018

© Copyright by Ioannis Branikas, 2018.

All rights reserved.

Abstract

This dissertation consists of three papers on the portfolio choices of retail and institutional investors.

The first chapter estimates the effects of product market advertising on household portfolio choice. Product market advertising by raising the awareness of a company's brand is thought to increase the demand for a company's stock as well as its products. I construct a dataset of publicly traded sports sponsors in the US and develop an instrument for investor exposure to advertising via these sponsorships. I show that investors living in a city where local sports teams are sponsored by a given company, local or non-local, are more likely to purchase stocks in that company. The portfolio effects from sports sponsorship are large and suggest that advertising is more important than even local bias.

The second chapter, co-authored with Harrison Hong and Jiangmin Xu, revisits the local bias in household stock investments. Households hold undiversified stock portfolios of firms headquartered near their city of residence. Leading explanations such as the familiarity heuristic assign a causal role to proximity. However, households optimally locate in a city based on unobservables such as optimism about a city's prospects, which can be correlated with latent local-stock demand. Using location choice models, we develop an instrumented-Heckman (1977)-style correction for widely-used regressions to show that local bias is also driven by this selection. We then propose a decomposition of local bias into household priors about stocks, using location-choice model residuals, and familiarity, whereby confidence about stock-payoff signals rise with proximity.

The third chapter, also co-authored with Harrison Hong and Jiangmin Xu, estimates the costs of producing information in financial markets. Information in stock

markets is often produced by sell-side analysts working at banks, whose research is bought by institutional investors. We estimate the costs and benefits of this information by modeling the effect of analyst coverage on institutional stock portfolios. We quantify two distinct channels: coverage drawing managerial attention to stocks versus improving the precision of managerial signals. We can then calculate the marginal benefit of information as the elasticity of aggregate institutional demand for a stock to the banks allocating one analyst to that stock. Assuming convex cost functions for coverage, we can then measure the marginal cost producing information for each stock.

Acknowledgements

I am very much grateful to the members of my dissertation committee, Harrison Hong, Jakub Kastl and Motohiro Yogo, for their guidance and support.

First and foremost, to Harrison Hong, with whom two of the three chapters in this dissertation are co-authored. Harrison showed trust and confidence in me during a critical stage of my PhD studies. His mentorship has been a true blessing.

Jakub Kastl, the head of my committee, patiently advised and encouraged my research ventures throughout the PhD program. Motohiro Yogo provided valuable feedback. His work on the portfolio demand of institutions was a great source of inspiration.

I also thank my co-author and friend, Jiangmin Xu, for bringing composure and coherence in our projects. My education in the methods of Industrial Organization is the result of the teachings of Jakub Kastl, Jan De Loecker and Eduardo Morales. I also benefited a lot from discussions with Bo Honore and Kirill Evdokimov. My good friend and classmate, Constantinos Kalfarentzos, is responsible for the significant enhancement of my programming skills.

The work in this dissertation had heavy computational requirements. Matthew Parker's technical assistance was a big help. I also thank Oscar Torres-Reyna and Ofira Schwartz-Soicher for allowing me to use extensively the computers in the DSS lab.

The first chapter of this thesis could not have been written without the marketing data that were obtained thanks to the Interlibrary Loan Services in Firestone. I also thank Bobray Bordelon, Joann Donatiello and Elana Broch for helping me access additional data that I needed.

For financial support during the PhD program, I thank the Graduate School, the Department of Economics and the Seeger Center for Hellenic Studies at Princeton University.

For helping me deal with paperwork and their assistance in the organization of the PhD Workshops in Finance and Industrial Organization, I thank Laura Hedden, Jessica O'Leary, Karen Neukirchen, and Gabriella Herrera.

Last, but definitely not least, I thank my classmates and friends. Dmitry Mukhin is the mature mind of our group. Jakob Schlockerman is the reduced-form labor economics expert. Gaurav Sabharwal is the young philosopher. Gabriel Buchbinder is the dear smart friend.

The dissertation is dedicated to my mother, Theoni. My efforts are for her.

To my mother

Contents

1 Advertising Exposure and Portfolio Choice **1**

1.1 Introduction 1

1.2 Data 8

1.2.1 Sports Sponsors 8

1.2.2 Household Stock Holdings and Demographics 10

1.2.3 Stock Characteristics 14

1.3 Empirical Approach 16

1.3.1 Portfolio Choice 16

1.3.2 Instrument of Sports Sponsorship 18

1.3.3 Control Function Approach 21

1.4 Estimation 23

1.4.1 Controlling for Observable Fundamentals 23

1.4.2 Controlling for Unobservable Fundamentals 26

1.5 Assessing the Advertising Over-Investment 28

1.5.1 Portfolio Trades 28

1.5.2 Portfolio Returns 31

1.5.3 Directly vs. Non-Directly Recognizable Stocks 33

1.6 Unbundling Familiarity 35

1.6.1 Portfolio Demand 35

1.6.2 Portfolio Diversification 37

1.7 Conclusion . . . 39

2 Location Choice, Portfolio Choice **63**

2.1 Introduction . . . 63

2.2 Accounting for Location Choice in Reduced Form Portfolio Choice Regressions . . . 68

2.2.1 Location Choice . . . 69

2.2.2 Reduced-Form Portfolio Choice Regressions . . . 71

2.2.3 Selection Correction . . . 72

2.3 Data . . . 75

2.3.1 MSA Demographics . . . 75

2.3.2 Household Demographics . . . 76

2.3.3 Geographical Distribution of Household Stock Holdings . . . 77

2.3.4 Stock Financial Characteristics . . . 79

2.3.5 Summary Statistics on Local Bias of Stock Portfolios . . . 80

2.4 Identification . . . 81

2.4.1 Exclusion Restriction . . . 81

2.4.2 Specification of the Correction Function . . . 81

2.5 Estimation . . . 84

2.5.1 Location Choice Results . . . 84

2.5.2 Portfolio Choice Results . . . 86

2.5.3 Robustness . . . 89

2.6 A Decomposition of Local Bias . . . 90

2.6.1 Model . . . 90

2.6.2 Estimation . . . 94

2.6.3 Model Fit . . . 95

2.7 Conclusion . . . 96

3 Estimates of Costs and Benefits of Information in Stock Markets 119

3.1 Introduction . . . 119

3.2 Model . . . 124

3.2.1 Consideration Set Formation . . . 124

3.2.2 Portfolio Choice . . . 126

3.2.3 Fund Returns from Analyst Coverage . . . 129

3.3 Data . . . 130

3.3.1 Mutual Fund Stock Holdings . . . 130

3.3.2 Stock Characteristics . . . 132

3.4 Maximum Likelihood Estimation . . . 133

3.4.1 Log Likelihood . . . 133

3.4.2 Coefficient Estimates . . . 134

3.4.3 Model Fit . . . 136

3.4.4 Analyst Coverage Economic Effect . . . 137

3.5 Analyst Coverage Instrument . . . 139

3.5.1 Mergers of Brokerage Firms . . . 139

3.5.2 DID Estimation . . . 141

3.5.3 Control Function Estimation . . . 142

3.6 Robustness . . . 144

3.6.1 Analyst EPS Forecasts and Stock Recommendations . . . 144

3.6.2 Financial Characteristics in the Extensive Margin . . . 149

3.7 Mutual Fund Returns from Analyst Coverage . . . 151

3.8 Estimates of the Marginal Cost of Producing Information . . . 153

3.8.1 Banks' Stock Coverage Decisions . . . 153

3.8.2 Correlations with Stock Characteristics . . . 155

3.9 Conclusion . . . 155

Bibliography 188

Chapter 1

Advertising Exposure and Portfolio Choice: Estimates Based on Sports Sponsorships*

1.1. Introduction

Where do households get investment ideas? Addressing this question is essential for understanding the portfolio choices of individuals, whether these refer to stocks, mutual funds or products of financial innovation.[1] *Product market advertising* is a widely discussed source for investment ideas, but with little empirical evidence. The premise is that the awareness of a firm's securities can be raised through the promotion of the firm's products. Thus, besides increasing product sales, which is its

*I am indebted to my advisors Harrison Hong, Jakub Kastl and Motohiro Yogo. For comments and discussions, I also thank Bo Honore, Atif Mian, Jiangmin Xu, Maryam Farboodi, David Schoenherr, Michal Kolesar, Isaiah Andrews, Mikkel Plagborg-Moller, Nick Buchholz, Dmitry Mukhin, Giulia Brancaccio, Gaurav Sabharwal, Shoumitro Chatterjee, Julia Fonseca, Terry Moon, Liyu Dou, Donghwa Shin, Jakob Schlockermann, Gabriel Buchbinder, David Zhang, Spyros Pagratis, Constantinos Kalfarentzos, Charis Katsiardis and participants at the Civitas Finance Seminar and the Finance, Industrial Organization and Econometrics Workshops at Princeton University.

[1]See, e.g., Shiller and Pound (1989), Hong, Kubik, and Stein (2004), Barber and Odean (2008), and Bailey, Kumar, and Ng (2011).

main objective, advertising can potentially also increase the retail investors' demand for a stock through a spillover effect (e.g., Grullon, Kanatas, and Weston (2004), Lou (2014))[2].

This paper seeks to contribute to the identification of the product advertising effect on the household portfolio, by utilizing account data from a large national discount broker in the United States in conjunction with a newly constructed dataset of publicly traded sponsors of teams in the four major league sports (i.e., NFL, NBA, MLB and NHL). In particular, the aim is to use micro-level retail investment data in order to study how product market advertising interacts with the acknowledged presence of locally biased equity preferences.[3] Most interestingly, the paper's final goal is an investigation of whether geographically distant firms can actually make households' stockholdings *less local* by sponsoring a sports team in the city where they live. The extent to which local firms can also attract further retail investment in the city where they are based, if they sponsor a local team, is examined too.

The biggest research challenge for Household Finance studies in advertising is the general unavailability of data on the exposure of investors to commercials. For instance, the advertising expenses from the account balance sheets of firms in Compustat (which are used in many papers) are too crude to know which population in which local market is exposed to advertising. Additionally, in the US, there is lack of data on individuals' stock-holdings to begin with. As a result, the study of the typical American retail investor relies almost exclusively on the account data of the large discount broker that this paper uses.[4] Unfortunately, due to their age, these data cannot be merged with the more recent local advertising data from Nielsen or

[2]See also Fehle, Tsyplakov, and Zdorovtsov (2005), Chemmanur and Yan (2009), Keloharju, Knüpfer, and Linnainmaa (2012).

[3]Particularly for retail stock investment, the Household Finance literature has established the renowned channel of *local bias*, according to which equity portfolios of households are highly concentrated on firms with headquarters near their residence. See, e.g., Grinblatt and Keloharju (2001), Huberman (2001), Ivković and Weisbenner (2005), Massa and Simonov (2006), Feng and Seasholes (2008).

[4]See Barber and Odean (2013) for a literature review.

Kantar Media.[5]

But even with data on the local marketing activities, different forms of advertising (e.g., TV or radio commercials, street or mail advertising, etc.) have in general different effects. Notably, recent papers that study the household health insurance choice document a decreasing efficiency of media advertising (e.g., Aizawa and Kim (2015), Shapiro (2017)). This finding is not surprising, since in principle households do not enjoy watching commercials and try to skip them if they can.[6]

To overcome the above issues, in this paper, I measure advertising exposure with sports sponsorship. A well-accepted notion in Marketing is that sports teams offer the most suitable platforms in which firms can increase their visibility (e.g., Carter (1996), Ukman (2017)). Sports sponsorship is a form of advertising that is perceived to be less obtrusive by individuals, since it associates brands with the cause of supporting their team. According to a nationwide poll conducted by John Hancock Financial Services in 1993, up to 64% of a city's households "would think more favorably" of a company that sponsors their local team. In fact, the anticipated benefits from sports sponsorship are such that teams usually provide the basis for an integrated advertising campaign in a city. That is, rather than being a substitute, sports sponsorship can complement any other form of advertising.[7] As a result, even in the absence of advertising data in local markets, sports sponsorship can be used as a *sufficient statistic* for the increased overall marketing efficiency of a firm in a city.

Based on the above remark, I employ a firm-city-specific indicator variable that equals one if the respective firm sponsors a team in the corresponding city. To construct this variable, I digitize the Sports Sponsor FactBook of Team Marketing Report, which allows me to see the sports marketing activities of primary sports sponsors in all cities. I am then able to compare the households' portfolio positions in

[5]See, e.g., Madsen and Niessner (2016), Focke, Ruenzi, and Ungeheuer (2016), Hillert, Kunzmann, and Ruenzi (2017).

[6]See, e.g., Speck and Elliott (1997), Cho and Cheon (2004) .

[7]See, e.g., Meenaghan (2001), Carrillat and d'Astous (2012), Olson and Thjømøe (2009).

their cities' team sponsors with their stockholdings in other firms. Since most sports sponsors are large firms, to make the comparison effective, I contrast them to other members of Russell 1000 (i.e., an index with the largest 1000 public companies based on market capitalization).

To provide a first perspective, Table 1.1 depicts the average allocation of the household portfolio on firms based on (i) how far away their headquarters are from the household residence and (ii) whether they sponsor a professional sports team in the city where the household lives. Column 3 in Panel A shows the standard local bias. On average, 21.1% of the household portfolio is invested in stocks whose headquarters are at most 100 miles away from the household residence. That portion is 14.7% in excess of the share according to the market (value-weighted) portfolio. At the same time, Row 3 shows that 11.7% of the portfolio is on average allocated on the team sponsors of a household's city. In fact, relative to the CAPM benchmark, there is an over-investment of 5.4%. Importantly, as Column 1 demonstrates, *the over-investment in sports sponsors cannot be absorbed by local bias.* That is, 24% of the advertising over-investment (i.e., 1.3% out of the 5.4%) is attributed to stocks with headquarters more than 100 miles away from the household residence. Even if the locality threshold is extended to 250 miles (in Panel B), stocks based further away continue to comprise 17% of the bias towards stocks that advertise (i.e., 0.9% out of 5.4%).

To capture the high degree of portfolio sparsity that retail investors exhibit (owning on average two to three stocks), I use a Tobit regression framework of household portfolio weights on firm sports sponsorship, geographical proximity, stock characteristics, household and city demographics. This empirical setup, which is very similar to the contemporaneous work by Branikas, Hong, and Xu (2017), is suitable for capturing the sports sponsorship effect on both the intensive and extensive margin of

household portfolios.[8]

I propose an instrument for the investor exposure to sports sponsorship that is based on the interaction of sports team presence in a city with a firm's propensity to advertise. The identification concept is analogous to a difference-in-difference setup (e.g., Duflo (2001), Nunn and Qian (2014)) in the firm-city cross-section. To limit the impact of city unobservables, I only focus on households living in 71 Metropolitan Statistical Areas (MSAs) which were in the *market for a sports team*, i.e., in MSAs in which professional sports teams were located or were considering relocating. To measure a firm's advertising propensity, I use the product of its size with the total advertising expenditure in its industry. As a result, the treatment refers to the sports sponsorships in cities with sports teams, made by large firms in industries with high advertising expenses.

In the presence of an extensive list of controls (which includes, among others, the Fama and French (2015) factors, geographical and professional industrial proximity as well as the individual terms of sports team presence and firm advertising propensity), it is unlikely that household investments will depend on the treatment effect for any other reason than the actual sports sponsorship. Perhaps, one story could have been that households in cities with sports teams might better assess firms with certain features (e.g., high advertising propensity). But such a scenario seems improbable, since the demographics in the selected 71 MSAs are comparable.[9]

I follow a control function approach (e.g., Petrin and Train (2010), Wooldridge (2014)) to predict the probabilities with which the firms in the investment universe would sponsor sports teams in the selected cities (based on the city sports team presence - firm advertising propensity interaction). Once I estimate these probabilities, I incorporate them as additional controls in the censored portfolio weights regres-

[8] Branikas, Hong, and Xu (2017) show that households' local bias can be explained by latent factors that drive their location choice.

[9] Grinblatt and Keloharju (2001) find language effects on the portfolio choice in Finland. Yet, in the US, such effects seem impossible.

sion, in order to correct, in the style of Heckman (1978), for the observed firm sports sponsorship in a given city.

I find that sports sponsorship increases a household's portfolio weight on a stock by about 50% relative to the average (i.e., about 4 bps when the average portfolio weight on a Russell 1000 stock - including the many zero positions - is 8 bps). Additionally, in terms of trading, sports sponsorship also increases a household's monthly stock buys by 36% relative to the average. However, there is no statistically significant decrease in the household's monthly stock sells. This suggests that product advertising is more relevant for the purchase decisions of retail investors. Notably, the positive effect on the household portfolio choice holds for stocks whose name can be directly recognized from the name of the product that is advertised.

In principle, the over-investment in stocks that advertise, can be explained by the same classes of models which interpret local bias. The local bias theories revolve around the causal role of geographical proximity. In rational models, investors know or can acquire more information from firms based close to them.[10] This implies that local stock-holdings are expected to yield higher returns than distant ones. In contrast, in behavioral models, proximity draws investors' limited attention and breeds a familiarity bias which is independent of portfolio returns.[11] Seasholes and Zhu (2010), however, show that the local bias portfolio under-diversification leads to underperformance. In line with this finding, I document an underperformance for the portfolio under-diversification that sports sponsorship causes. Moreover, given the advertising effect on stock choice that I identify, geographically distant firms can attract more retail investment by sponsoring a city's sports team. The interpretations of attention or salience seem thus to be more consistent with the data.[12]

[10] See, e.g., Merton (1987), Wang (1993), Coval and Moskowitz (2001), Van Nieuwerburgh and Veldkamp (2009).

[11] See, e.g., Sims (2003), Peng and Xiong (2006), Agarwal, Driscoll, Gabaix, and Laibson (2009), Gennaioli and Shleifer (2010).

[12] Along similar lines, Fang, Peress, and Zheng (2014) show an attention effect on the portfolio choice of mutual fund managers from the coverage of stocks in the media.

I further analyze how local bias and product advertising interact with each other in the household portfolio demand and the consequent portfolio under-diversification. As in Table 1.1, I classify stocks into four classes, based on whether they are local or distant (given their headquarters' proximity from a household's residence) and whether they sponsor a team in a household's city or not. I find that households have the strongest portfolio preferences for stocks that are local *and* sponsors of their local teams. By sports sponsoring their headquarters' city, local firms can increase their share in the household portfolio by about 20 bps (i.e., 250% relative to the average). Therefore, product advertising largely magnifies the local bias.

Importantly, investors living in a city where local teams are sponsored by a given non-local firm are far more likely to purchase stocks in that firm than in other distant firms which do not sports sponsor their city. The expected increase in the portfolio share amounts approximately to 40% relative to the average. In fact, the portfolio effects that the sports sponsorships of distant firms generate in a city can outweigh the local bias for stocks which do *not* sponsor a local team. Whether this happens or not depends on how far away the firms' headquarters are located. In particular, if locality is defined based on a 250-mile headquarters' distance from a household's residence, the sports sponsorship effect of distant stocks is found to be stronger than the investment bias for local firms that do not sports sponsor. As a result, the investment channel of product market advertising is found to matter more than the local bias.

The organization of the paper is as follows: In Section 2, I describe the data on the publicly traded primary sports sponsors and the retail investors' stockholdings. In Section 3, I discuss the empirical model and the identification strategy. In Section 4, I present the main empirical findings for the sports advertising effect on the household portfolio choice. In Section 5, I assess the over-investment in sports sponsorship, by estimating its impact on the household portfolio trades and returns as well as the

relevancy of the direct recognizability of a stock from the names of its products. In Section 6, I study how the familiarity from sports sponsorship interacts with local bias. In Section 7, I conclude.

1.2. Data

1.2.1. Sports Sponsors

Data on the sports advertising of publicly listed firms are extracted from the 1993, 1994, 1995 and 1996 issues of Team Marketing Report's Sports Sponsor FactBook. Every issue refers to the contemporary sports season and contains a detailed description of the sports marketing activities of all sports sponsors at the brand level.[13] I only focus on the firms that Team Marketing Report characterized as *Primary Sports Sponsors* (i.e., "the most significant and active sports sponsors"). I use the CRSP US Stock Database to identify which of these firms were publicly traded at that time and find their historic PERMNO and CUSIP codes.

I restrict my attention to marketing activities in Major League Baseball (MLB), the National Basketball Association (NBA), the National Football League (NFL) and the National Hockey League (NHL). I record all the sports teams that the firms sponsored in the four professional leagues. I use the websites of Wikipedia and Sports Reference to find the stadiums in which the teams played their home games in each season.[14] The FactBook uses three main labels to describe the advertising of firms, namely (i) Sponsorship, (ii) Advertising and (iii) In-Stadium Signage. For a given sports sponsorship, these labels are not mutually exclusive. Although I have recorded these details, in the empirical analysis that follows, I am only interested in whether

[13]For example, the 1994 Sports FactBook was published on October 15th 1993 and referred to the 1993-1994 sports season.

[14]I account for the fact that, in some seasons, a few teams (e.g., Boston Celtics) used more than one stadium (e.g., the TD Garden and the Hartford Civic Center).

a publicly traded firm was marketed in a given Metropolitan Statistical Area (MSA) during a given sports season. To this end, I simply employ a firm-city-specific dummy variable, namely $SportsAd$, which equals to 1 if a firm sponsored a sports team in a given city.

In the FactBook, the sports marketing activities of firms are described at the brand level. Thus, in the process of the above aggregation, I am able to see how much overlap there is between the name of an advertised product and the name of its producer. In cases in which the name of a brand is very different, I use the websites of Youtube and Adland to see if the company name or logo was displayed on the TV commercials that were aired in the 90's. With this criterion in mind, when I construct the dummy variable $SportsAd$, I assign the value 1 only if the company could be directly recognized from the sponsorship. I use the same filter also for public holding companies which were parents of primary sponsors.[15]

Table 1.2 describes the distribution of the directly-recognizable publicly traded primary sports sponsors over time.[16] Their average number in a sports season is 170. About 70% of these firms belong, on average, in the Russell 1000 Index (which includes the largest 1000 stocks based on market capitalization). From the remaining 30%, 10% corresponds to foreign firms (with US headquarters) traded via an American depository receipt (ADR).

When assigning the publicly listed sponsors in one of the 17 Fama-French industry portfolios, most of them (i.e., 17% on average) belong to the portfolio of retail stores (Rtail). 14% are in the financial portfolio (Finan), while another 14% are in the "other or services" portfolio (OthSvc). Similarly, about 10% are in the portfolio of

[15]For example, drug commercials (e.g., Advil) typically did not display the name of their manufacturer (e.g., American Home Products). Therefore, in most cases, pharmaceutical stocks were not directly recognizable from their sponsorship. Similarly, it was difficult for someone to think of Philip Morris while watching a product commercial of General Foods USA, which was one of the company's subsidiaries at that time.

[16]The distribution of the non-directly-recognizable publicly traded primary sports sponsors is given in Table 1.10.

food (Food), while another 10% are in the portfolio of transportation (Trans). About 9% are in the portfolio oil and petroleum products (Oil).

Furthermore, during a given sports season, a publicly traded sports sponsor advertises on average in 4 Metropolitan Statistical Areas (MSAs), with the median number being equal to 2 or 3. Therefore, it is not the case that sports sponsors advertise only in the cities where they are headquartered.

1.2.2. Household Stock Holdings and Demographics

Data on household stock holdings are drawn from the database of a large national discount broker. See Barber and Odean (2000) for a detailed description. The dataset is an unbalanced panel for the period 1991-1996 and contains month-end account statements of 78,000 retail investors at the stock level (CUSIP code). As Ivković and Weisbenner (2005) report, the majority of accounts is non-retirement (e.g., cash or investment) and the few retirement accounts do not refer to 401(k) plans. Therefore, mechanical effects on the stock choice from the shares of an employer are limited.[17]

Not all investors have complete demographic information. In my analysis, I only keep households which have a non-missing address ZIP code and non-missing income, family size, age and gender of their head. Unfortunately, the data do not contain any information about the education, race and industrial sector of the household's head. To correct for that, I follow Korniotis and Kumar (2011) and extract from Census 1990 the education status (i.e., the probability of holding a B.A. or higher degree) and the racial profile (i.e., the probability of being White, Black, Hispanic, Asian or other) at the household's ZIP Code. Moreover, I use the distribution of the employed persons into industries at the ZIP Code level to measure the household's (expected) professional industrial proximity to a stock, in the style of Massa and Simonov (2006).[18]

[17] Employee stock ownership plans were not popular in the early 90's.

[18] For example, New Yorkers living in Upper East Side are expected to be familiar with stocks in

From the address ZIP codes and the geographic relationship files of the US Census Bureau, I identify the Metropolitan Statistical Areas (MSAs) in which households reside. To construct my identification design, I only focus on 71 MSAs which were in the "market for a sports team". On average, every year, that set includes about 38 MSAs in which professional sports teams were present and 33 MSAs which did not have any professional sports team, but whose local governments were trying to persuade franchise owners to move in. The latter subset of MSAs is based on the lists of cities provided by Euchner (1993) and Danielson (1997).

For the selected 71 MSAs, I collect annual demographic series from a variety of sources. From the Bureau of Economic Analysis (BEA), I draw the total income, while, from Bureau of Labor Statistics (BLS), I extract the population and unemployment rate. Additionally, following Bishop (2007), I construct a measure of median home values (for every MSA in every year of the sample) based on the time series of the house price index (HPI) of the Federal Housing Finance Agency (FHFA) and the median home values in Census 2000.[19]

The universe of stocks that I examine is the union of the publicly traded sports sponsors (described in the previous subsection) and the set of stocks that were ever members of Russell 1000 during the sample. Since 70% of the firms with sports marketing activities belong to Russell 1000, the remaining stocks in the index comprise an appropriate comparison group for the study of the advertising effect on portfolio choice (see Panel G of Table 1.3). The ever-membership requirement protects the identification from mechanical effects on the stock choice caused by the Russell's 1000 reconstitution. I also require stocks to have observable financial characteristics at a monthly frequency (described below). Retail investors who do not hold any stocks in the universe are omitted from the analysis.[20]

the financial sector, since many investment bankers reside there.

[19]Relying solely on the HPI would have limited the house price variation across MSAs, since the HPI of all MSAs has as base year the year 1995. I thank Selale Tuzel for this comment.

[20]Given the definition of the stock choice set, all other stocks can be viewed as outside assets.

In every month, the portfolio weights of a household are constructed by dividing the dollar value on a stock with the total dollar value of the household's holdings in the universe. Most households in the data have multiple accounts. When I derive the dollar value of an investor's position, I take into account all his accounts. Households whose portfolio value is less than 1,000 dollars are dropped from the analysis.

To study the impact of the sports sponsorships, I focus only on the monthly statements in the sports seasons for which I have advertisement data from the FactBook. Specifically, I start my sample period in September 1992 and end it in August 1996. All the monthly statements from the September of year t until the August of year $t+1$ are assigned to the sports season that ends in year $t+1$.

Applying these filters leads us to a sample of 18,421 different households and 1,407 different stocks for the whole period.[21] On average, in every month, there are 9,462 households choosing from 1,256 stocks with a complete list of financial characteristics. The geographical distribution of the households, along with the headquarters of the stocks and sports teams is depicted in Figure 1.1, on a map of latitude and longitude coordinates. Overall, the sample is dispersed enough to be representative of the US population. Moreover, in terms of the potential local bias, households are always located near the headquarters of some firms. MSAs without sports teams are also not that far away from MSAs with sports teams.

To assess the trading activity of the selected households, I calculate their sales and purchase turnover as Barber and Odean (2000). On average, the monthly sales turnover is 3.20%, while the monthly purchase turnover is 4.05%. In other words, the retail investors in the sample are not passive, since they buy 48.6% and sell 38.4% of their portfolio every year.[22]

The summary statistics of the household stock holdings are presented in Panel A

[21] Although the beginning and end of the households' time series is different, there is a decreasing trend in their number (from 14,280 households in September 1992 to 6,290 in August 1996).

[22] For all households and all stocks in the database, Barber and Odean (2000) document an average annual portfolio turnover of 75%.

of Table 1.3. The average portfolio value of a household is 38.44 thousand dollars, while the median is 13.30 thousand dollars. The standard deviation of the portfolio value is 143.36 thousand dollars. In line with the observed portfolio sparsity, the median portfolio weight is zero. On average, a household holds 2.73 stocks in a given a month. The median number of stocks that are held is 1.92, while the standard deviation is 2.78. The average portfolio weight of a household on a stock in the universe is 7.97 basis points (bps), while the standard deviation is 234.58 bps.[23]

Panel B depicts the summary statistics for the stock geographical proximity and sports advertisement in the investment universe. The average distance between the address ZIP code of a household and the address ZIP code of a stock's headquarters is 920.54 statute miles. The standard deviation is 604.01. On average 84% (93%) of the investment universe is at least 250 miles (100 miles) away from the household residence. Moreover, on average, 2% of stocks in the investment universe sponsor a sports team in the city where a household lives (taking into account also households in cities without sports teams).

Panel C of Table 1.3 presents the summary statistics of the household's demographics. The average income is 93.43 thousand dollars, while the median is 87.50 thousand dollars. Households are on average 54.04 years old, with the median age being equal to 50. Moreover, 89% of the retail investors are on average men. The average family size of a household is 2.47, while the median is 2.

Panel D of Table 1.3 describes the summary statistics of the MSAs in which the households live. On average 53% of the MSAs have a professional sports team. The population is on average 2.07 million, with a median equal to 2.66 million. The average income per capita in a MSA is 23.80 thousand dollars, with a standard deviation of 3.02. The median home value is on average 97.25 thousand dollars and has a standard deviation of 30.36. The unemployment rate is on average 5.17% and has a standard

[23]Conditioning on the household's stockholdings (i.e. looking only at the intensive margin of the household portfolio), the average portfolio weight on a stock 0.37, with the median being 0.23.

deviation of 1.43.

1.2.3. Stock Characteristics

Monthly data on stock prices and returns are drawn from CRSP, while firm accounting variables are collected from Compustat at a quarterly frequency. Stocks with missing observations are dropped from the sample. The list of financial variables that I use consists of the price, the market capitalization (Size), the book-to-market ratio (Book/Market), the turnover ratio (i.e., Turnover, defined as volume over number of shares outstanding), the momentum (i.e., Momentum, defined as the past annual return), the volatility (i.e., Volatility, defined as the standard deviation of monthly returns in the past year), profitability (i.e., Profitability, defined, as in Novy-Marx (2013), as the ratio of past annual gross profits to assets), the investment (i.e., Investment, defined as the past annual growth rate of assets) and the past annual sales (Sales). All these variables are constructed at a monthly frequency. I assume that a household's investment decision in month t is based on the stocks' price in that month and the above risk factors in month $t-1$.[24]

From Compustat, I also extract data on the firms' advertising expenses. To fill in some missing observations, I merge the Compustat data with the contemporary annual reports on the 200 Leading National Advertisers, provided by the Advertising Factbook of the Advertising Age magazine.[25] The industry advertising expenditure, in each of the 17 Fama-French industry portfolios, is calculated by summing up the advertising expenses of all industry participants in the Compustat universe (and not just the investment universe).

Since Compustat contains only the most recent headquarters' addresses of the stocks, a variety of sources is utilized to ensure that the headquarter information in

[24] As in Fama and French (1992), to make sure that firms' balance sheet information is known to investors, I match the accounting variables from the fiscal year $t-1$ with the stock prices from the July of year t until the June of year $t+1$.

[25] Historical copies were obtained and digitized from the LexisNexis Database.

the sample period is accurate.[26] In particular, the addresses of all sports sponsors are obtained from the Sports Sponsor FactBook of Team Marketing Report. For the other firms, I use the SEC's EDGAR, Comphist and a digitized sample from the Who Owns Whom Database.[27]

Panel E of Table 1.3, depicts the time-series average of the total advertisement expenditure in an industry. As expected, in terms of ranking, the industries of services (OthSvc), food (Food), consumption (Cnsum) and retail stores (Rtail) are first. The advertising expenditure of every firm in the investment universe is summarized in Panel F. In particular, the overall average is 219.27 million dollars, with the median being equal to 59.36 million dollars. The publicly traded primary sports sponsors are heavy spenders of advertising. Their average expenses on ads is 374.77 million dollars and their median equals 158.98 million dollars. The other firms in Russell 1000 spend, on average, about 60% less (i.e., 149.75 million dollars). Their median is about 1/5 of the sports sponsors' median (i.e., 35 million dollars).

The summary statistics of the financial characteristics of a stock in the universe, during the sample period, are depicted in Panel G of Table 1.3. The average price is 34.79 dollars, with a standard deviation of 24.37. The average size is about 3.5 billion dollars, while the median size is 1.3 billion dollars. Book-to-market is on average 6.07, with a standard deviation of 113.41. The average monthly turnover ratio is 10%, with a corresponding standard deviation of 0.13. Momentum is on average 10%, with a standard deviation of 0.60. Volatility is on average 8% and has a standard deviation of 0.05. Profitability is on average 0.33 and Investment 0.19. The average sales of a firm equal roughly 4 billion dollars, while the median 1.2 billion dollars.

[26]For example, Boeing Co. is nowadays headquartered in Chicago, but, during the 90's, it was based in Seattle. Pirinsky and Wang (2006) identify 118 firm relocations from 1992 to 1997 and Tuzel and Zhang (2017) about 300 from 1990 to 2005. Usually, firms that move are small and not members of Russell 1000.

[27]The SEC's EDGAR data are obtained from Bill McDonald's website at https://www3.nd.edu/~mcdonald/10-K_Headers/10-K_Headers.html. I thank Christo Pirinsky for letting me know about this source.

The panel also indicates the summary statistics of the financial characteristics of the publicly traded primary sports sponsors and the other firms in Russell 1000 separately. In terms of turnover, momentum, volatility, profitability and investment, the two groups of stocks are similar. In terms of market capitalization, firms which advertise are bigger. The average size of a publicly listed sports sponsor is about 10 billion dollars, with the median being equal to 3.6 billion dollars. On the other hand, the average size of the other firms in Russell 1000 is 2.5 billion, with the median being equal to 1.2 billion dollars. Although the mean and standard deviation of book-to-market differs between the groups, their median is similar.[28] Besides size, sports sponsors have also larger (past) sales. Their average and median equal 13.7 and 6.5 billion dollars respectively, while other stocks have a mean of 2.4 and a median of 1.1 billion dollars.

1.3. Empirical Approach

1.3.1. Portfolio Choice

To identify the impact of product market advertising on household portfolio choice, I setup a simple static empirical specification in the cross-section of households, cities and stocks. When I estimate the model, I run this specification for every month in the sample separately. I then focus on the interpretation of the average values of the estimated parameters. Therefore, there is no period subscript t in the discussion that follows.

In the retail investment data, there is a high degree of portfolio sparsity, in the sense that a household owns on average 2 to 3 different stocks. To capture that sparsity as well as the fact that households do not short, I assume that household i,

[28] The first and third quartile is also similar, (i.e. for the sports sponsors it is Q1=0.30 and Q3=1, while for the other firms in Russell 1000 it is Q1=0.24 and Q3=0.6.)

residing in city c, decides how much to invest in stock j according to a linear factor rule censored at zero, i.e.:[29]

$$w_{i,c,j} = (\alpha + \beta SportsAd_{c,j} + \boldsymbol{\gamma}\boldsymbol{X}_{i,c,j} + \epsilon_{i,c,j})^{+} \tag{1.1}$$

where $(\cdot)^{+} \equiv \max\{\cdot, 0\}$ captures household i's extensive and intensive margin. $SportsAd_{c,j}$ is a dummy variable that equals 1 if stock j sponsors a sports team in city c (where household i resides) and 0 otherwise. Its coefficient (β) and marginal effect are the main objects of interest from the identification.

$\boldsymbol{X}_{i,c,j}$ is a vector of observable fundamentals that are introduced as controls. A key element is the (log) distance between household i's residence and stock j's headquarters at the ZIP Code level, which captures the well-established phenomenon of local bias. Other observable fundamentals refer to stock j's financial characteristics, such as the Fama and French (2015) factors (i.e., size, book-to-market ratio, profitability, investment) as well as extra factors (i.e., turnover, momentum, volatility, industry fixed effects) from the stock demand literature (e.g., Gompers and Metrick (2001), Goetzmann and Kumar (2008)).

Following the Finance literature on advertising (e.g. Grullon, Kanatas, and Weston (2004), Lou (2014)), I also include as a factor the firm's past annual sales. Moreover, to account for the heterogeneity of preferences or beliefs across investors, I control for household i's demographics (e.g., income, the head's age and gender, family size, education, race, professional industrial proximity to a stock) and city c's demographics (e.g., population, income per capita, the median home value, unemployment rate).

The error term, $\epsilon_{i,c,j}$, can be interpreted as household i's latent demand for stock j. It encompasses household i's optimism or pessimism about stock j's prospects as

[29] There is a number of ways to microfound this censored regression (e.g., Brandt, Santa-Clara, and Valkanov (2009), Hjalmarsson and Manchev (2012), Gârleanu and Pedersen (2013), Hortaçsu and Kastl (2015)), Hong and Xu (2015), Koijen and Yogo (2016).

well as private information. In line with a Tobit model, I assume that, conditional on all observables, these idiosyncratic demand shocks are distributed according to the normal distribution.

In the spirit of Petersen (2009), in the estimation of the model, I use two-way clustered standard errors at the household and city level. On the one hand, the clustering at the household level accounts for the joint decision of an investor's portfolio shares across stocks. On the other hand, the city-clustering is consistent with the fact that investors in the same city choose from the same local stocks and face the same local news and economic conditions.

1.3.2. Instrument of Sports Sponsorship

The extensive list of controls in Equation (1.1) does not eliminate the possible presence of unobserved firm-city fundamentals that could be related to the observed sports sponsorship. Local firm reputation or tastes, in particular, might be correlated with where a firm offers a sports sponsorship. In that case, a part of the estimated advertising effect might be due to some unobserved catering of existing tastes in the area.

For example, in 1993, Allied Signal Inc. (of Morristown, N.J.) sponsored the Pittsburgh Penguins instead of the New York Islanders. One year ago (i.e., in 1992) it had relocated its research engineering department from New York City to Detroit, firing a few hundred New Yorkers. It is thus likely that, at that time, Pittsburghers liked the firm more than New Yorkers in the first place, independently of any product advertising. But if it were so, then that could have been why Allied Signal offered a sports sponsorship in the City of Bridges and not in the Big Apple.

To address the endogeneity problem, I construct a sports sponsorship instrument based on two sources of exogenous variation. The first source of variation is in the cross-section of cities and refers to (major league) *sports team presence*. The second

source of variation is in the cross-section of firms and refers to *advertising propensity*. The intuition for the use of these two variables is explained below.

The presence of sports teams in a city is a necessary condition for sports sponsorship. In the US, on average, only about 38 cities have a team in the four major sports leagues every year.[30] At the same time, there are about 33 cities which do not have a professional sports team, but are trying to get one.[31] In terms of demographics, the former cities are not that different from the latter cities. In fact, they can be matched with them. In a given matched pair of cities (e.g., Salt Lake City vs. Albuquerque), sports team's presence, and thus sponsorship exposure, is determined by random factors, such as more favorable stadium deals, availability of better facilities and personal preferences of franchise owners.[32] Importantly, sports sponsorship is only a small fraction of a team's annual revenue (i.e. about 8% on average), which franchises can get in any of these cities.[33]

My proxy for the advertising propensity of a firm is given by the product of the firm's size (e.g., market capitalization) and the total advertising expenditure in the industry of the firm. Both a firm's size and industry are natural predictors of a firm's advertising expenses. Larger firms have the resources to spend more on advertising. And firms that are in industries with high advertising expenses (e.g., retail stores or food sector) need to advertise more to be competitive. Firm size times industry advertising expenses thus captures the tendency of big firms in industries with big advertising expenses (e.g., Pepsico Inc.) to advertise more and thus be affected more

[30] At least one professional sports team, to be exact. My analysis is not sports-specific and looks at the extensive margin of firm advertising in a city (captured by my firm-city-specific indicator variable).

[31] In the US, the professional sports teams are in limited supply. The leagues refuse to add more teams, so that teams can extract benefits from local governments. Given the opportunity and based on their demographics, the selected cities with no sports team (which are obtained from sports economics sources) were in position to address the requests of franchise owners.

[32] The Rams recently relocated from St. Louis to Los Angeles, in order to play in the Inglewood Stadium. The Rams had been playing in St. Louis since 1995, when Georgia Frontiere, the team's owner back then, brought them from Los Angeles to her home town.

[33] For their revenues, sports franchises rely mostly on TV rights and gate receipts.

by the presence of a sports team in a city.

To instrument the observed sports sponsorship of a firm in a city based on the two above sources of variation, I use the *interaction* of the city's sports team presence and the firm's advertising propensity. Importantly, I also include both these two individual terms as additional controls in the vector $\boldsymbol{X}_{i,c,j}$ of Equation (1.1). That is, the presence of sports teams in a city is perceived to be another city demographic variable, while a firm's advertising propensity is perceived to be another stock financial factor. In that way, the identification relies only on the interaction term being *excluded* from the portfolio equations and operating only through the observed sports sponsorship (conditional on the controls).

The pursued identification strategy is very similar to a difference-in-differences estimation (e.g., Duflo (2001), Nunn and Qian (2014)). In a "diff-in-diff" setup, the reduced-form estimation would have simply compared the difference in the investments of households in cities with sports teams relative to the investments of households in cities without sports teams, in stocks with high advertising propensity relative to stocks with low advertising propensity. Yet, here, the focus is on the effect of the actual sports sponsorship of a firm in a given city. The instrument switches that effect on, since a big firm in an industry with big advertising expenses is more likely to sports sponsor a city with a team.

A priori, there seems to be no obvious reason for which the proposed instrument might be affecting household stock investment through a channel different than sports sponsorship. A potential story according to which the households in the cities with sports teams can better assess firms with certain features (e.g., high advertising propensity) is unlikely to be true, since the demographics of the cities which are selected as well as the retail investors in the database who live there are comparable. Moreover, besides the Fama and French (2015) factors, the observable fundamentals for which I control in Equation (1.1) contain the household-stock geographical

and professional industrial proximity, i.e., the household-stock-specific variables that matter most.

1.3.3. Control Function Approach

Based on the previous discussion, I setup a first-stage equation that predicts firm j's sports sponsorship in city c based on the proposed instrument. The sports sponsorship specification that I use is static and refers only to the cross-section of firms and cities. For every sports season, I estimate the model separately and compute average annual coefficients and economic effects. In particular, the first-stage equation is as follows:

$$SportsAd_{c,j} = \mathbf{1}\left[\kappa + \lambda ZAd_{c,j} + Controls_{c,j} + \omega_{c,j} \geq 0\right] \tag{1.2}$$

where $ZAd_{c,j} \equiv Sportsteam_c \times (LogSize_j \times LogInduAD_j)$ is the instrument of firm j's advertising exposure in city c. In more detail, $Sportsteam_c$ is a dummy variable equal to 1 if city c has a sports team, $LogSize_j$ is the market capitalization of firm j (in logs), while $LogInduAD_j$ denotes the total advertising expenditure in firm j's industry (also in logs). $LogSize_j \times LogInduAD_j$ is the measure of firm j's advertising propensity.

The controls in Equation (1.2) reflect more aspects of firm j's advertising choice in a given city. They include firm j's distance from city c, firm j's size, the total advertising expenditure in firm j's industry and city c's demographics (i.e. population, income per capita, median home values, unemployment rate). As for the error term $\omega_{c,j}$, it captures unobservable factors that influence firm j's sports advertising in city c. Conditional on all observables, these firm-city idiosyncratic shocks are assumed to be independently and identically distributed according to the standard normal distribution. To also account for some additional firm and city unobserved heterogeneity, in the estimation of Equation (1.2), standard errors are clustered at the firm and city

level.‘

With Equation (1.2) in hand, any potential correlation between the household latent demand for a stock ($\epsilon_{i,c,j}$) and the stock's sports sponsorship in the household's residence city ($SportsAd_{c,j}$) is attributed to the correlation between the former and the unobservable advertising factors that the stock faces in the city ($\omega_{c,j}$). Using a control function approach (e.g., Petrin and Train (2010), Wooldridge (2014)), it is possible to account for the correlation of the two errors by adding functions of the probability with which the firm is predicted to sports sponsor the household's city as additional explanatory variables in Equation (1.1). These probabilities are predicted following the first-stage estimation of Equation (1.2).

More clearly, letting $V_{c,j} \equiv \kappa + \lambda ZAd_{c,j} + Controls_{c,j}$ be the observable criterion with which firm j sports sponsors city c, the latent demand of household i, in city c, for stock j can be decomposed as follows:

$$\epsilon_{i,c,j} = \mathbb{E}\left(\epsilon_{i,c,j} \,|\omega_{c,j}, V_{c,j}\right) + \tilde{\epsilon}_{i,c,j} \tag{1.3}$$

where $\tilde{\epsilon}_{i,c,j}$ is an idiosyncratic household-city-stock specific investment error, which is, by construction, independent of firm j's sports sponsorship in city c. That is, $\tilde{\epsilon}_{i,c,j}$ is mean independent of all observables (and assumed to be i.i.d according to the standard normal distribution). As for the conditional expectation of the original latent demand, it can be written, in the spirit of Heckman (1978), as follows:[34]

[34]In a nutshell, the instrumentation of $SportsAd_{c,j}$ can be viewed as a generalized selection correction. For instance, if $SportsAd_{c,j} = 1$, then it is expected that the unobservable firm-city advertising factors are high enough for firm j to sports sponsor city c. How much these "favorable" unobservable advertising factors correlate with the household's portfolio latent demand depends on their joint distribution. In any case, the correlation will be a function of the observable advertising criterion of firm j for city c, which is estimated following the first-stage. Equivalently, by monotonicity, the conditional expectation of the household latent demand is a function of the respective (predicted)

$$\mathbb{E}\left(\epsilon_{i,c,j}\left|V_{c,j},\omega_{c,j}\right.\right)=\underbrace{SportsAd_{c,j}\cdot\Psi_1\left(p_{c,j}\right)+\left(1-SportsAd_{c,j}\right)\cdot\Psi_0\left(p_{c,j}\right)}_{\Psi(SportsAd_{c,j},p_{c,j})} \tag{1.4}$$

where $p_{c,j}$ is the probability that firm j sponsors a sports team in city c and $\Psi_0(\cdot)$, $\Psi_1(\cdot)$ are unknown correction functions.[35] Combining Equations (1.1), (1.3) and (1.4) yields that the equation to be taken to the data is:

$$w_{i,c,j}=\left(\alpha+\beta SportsAd_{c,j}+\boldsymbol{\gamma X}_{i,c,j}+\Psi\left(SportsAd_{c,j},p_{c,j}\right)+\tilde{\epsilon}_{i,c,j}\right)^{+} \tag{1.5}$$

In the estimation of Equation (1.5), I approximate the correction functions, $\Psi_0(\cdot)$, $\Psi_1(\cdot)$, with polynomials of second or third order.

1.4. Estimation

1.4.1. Controlling for Observable Fundamentals

First, I discuss the estimation results from Tobit portfolio weight regressions which control for observable fundamentals, but do not correct for the endogeneity of *SportsAd*. The results are presented in Table 1.4. They refer to the average monthly coefficient estimates of Equation (1.1) and the respective average t-statistics based on clustering standard errors at the household and city level.

sports sponsorship probability, i.e.:

$$\begin{aligned}\mathbb{E}\left(\epsilon_{i,c,j}\left|\omega_{c,j}\geq-V_{c,j},V_{c,j}\right.\right) &= \int_{-\infty}^{+\infty}\int_{-V_{c,j}}^{+\infty}\frac{\epsilon_{i,c,j}dF\left(\epsilon_{i,c,j},\omega_{c,j}\left|V_{c,j}\right.\right)}{\mathbb{P}\left(\omega_{c,j}\geq-V_{c,j}\left|V_{c,j}\right.\right)}\\ &\equiv \psi_1\left(V_{c,j}\right)=\Psi_1\left(p_{c,j}\right)\end{aligned}$$

Analogously, if $SportsAd_{c,j}=0$, then $E(\epsilon_{i,c,j}|\omega_{c,j}<-V_{c,j},V_{c,j})=\Psi_0(p_{c,j})$, where $\Psi_0(\cdot)$ is another correction function (different than $\Psi_1(\cdot)$).

[35] The correction functions are zero only if $\epsilon_{i,c,j}$ and $\omega_{c,j}$ are independent.

As reported in Column 1 of Panel A, in the absence of household demographics, stock characteristics and city demographics, but with the inclusion of the household-stock distance at the ZIP Code level in logs, the coefficient of $SportsAd$ is 0.682 and has a t-statistic of 14.81. In Column 2, which incorporates household demographics (i.e., $LogIncome$, $LogAge$, $LogFamSize$, etc.), the coefficient of sports sponsorship remains virtually unchanged, being equal to 0.680, with a t-statistic of 15.02. In Column 3, in which stock financial characteristics (i.e., $LogSize$, $Book/Market$, etc.) are also introduced, the coefficient of sports advertising drops approximately by 57%, becoming equal to 0.294, with a t-statistic of 7.06. Further adding city demographics (i.e., $LogPop$, $LogIncPerCap$, etc.) in Column 4 results in almost the same coefficient estimate as before (i.e., 0.300) and a t-statistic of 7.69.

The economic effect of sports sponsorship on the portfolio weight is defined as the marginal effect of a discrete change in $SportsAd$ from 0 to 1 and is calculated at the mean values of the controls. It is also estimated for every month separately. Its average monthly value is presented in Panel B of Table 1.4. Column 1 shows that, with only distance as a control, sports advertising impacts the portfolio weight by 32.46 bps. This figure corresponds to 407.3% of the average portfolio weight on a stock in the investment universe, when taking into account the extensive margin. With household demographics, in Column 2, the economic effect is very slightly decreased (as is the coefficient) to 32 bps or 401.5% of the mean portfolio weight. In Column 3, which also accounts for stock financial characteristics, the economic effect is decreased by approximately 85%. Specifically, it equals 4.79 bps or 60.1% of the mean portfolio weight. Introducing city demographics in Column 4, slightly raises the sports advertising impact to 4.92 bps or 61.7% of the mean.

As for the estimated coefficients of the controls, $LogDist$ always enters negatively and with a strongly statistically significant coefficient, which is in line with the households' local bias. The average value of its coefficient estimate does not change much

across the four columns. However, depending on controls, the implied economic effect of distance ranges from 2.31 bps to 1.33 bps (or equivalently from -29% to -16.7% of the mean).

Turning to the household demographics, $LogIncome$, $LogAge$ and $Male$ have positive and statistically significant coefficients. On the other hand, the coefficient of $LogFamSize$ is negative and statistically insignificant. The coefficients of the racial profile variables (i.e., $Black$, $Hispanic$, $AsianOther$) are also not statistically significant. Moreover, a household's expected education and professional industrial proximity, i.e. $Educ$ and $ProfProxim$, enter its investment equation with positive and statistically significant coefficients.

Among the estimated coefficients of the stock financial characteristics, the ones of $LogSize$, $Turnover$, $Volatility$, $LogSales$ and $LogSize \times LogInduAD$ (which measures the firm's advertising propensity) are positive and statistically significant. $Momentum$ and $Investment$ enter the household investment equation with negative and statistically significant coefficients. The estimated coefficients of $Book/Market$ and $Profitability$ are close to zero and statistically insignificant.

The coefficients of the MSA demographics are presented in Column 4. $LogPop$, and $LogIncPerCap$, have coefficients which are negative and statistically significant. The coefficient of $LogMedHome$ is also a statistically significant, but positive. The coefficient of $UnempRate$ is very small and statistically insignificant. $Sportsteam$ has a positive and statistically significant coefficient.

Based on the estimates from these regressions, even in the presence of all the observable fundamentals, the sports marketing activities of a publicly traded firm in a city are associated with a strong positive effect on that city's household portfolios (equal to 61.7% of the average household portfolio weight on a stock). In fact, the magnitude of that effect is stronger than the one from a stock's geographical proximity (which is 16.7% of the mean).

1.4.2. Controlling for Unobservable Fundamentals

First Stage

The results from the Probit regressions of firm sports sponsorships in cities, which Equation (1.2) describes, are presented in Table 1.5. The table depicts the average coefficient estimates and the average t-statistics from separate year-by-year estimations.

As reported in Panel A, the coefficient of the instrument, ZAd, is positive and strongly statistically significant, regardless of the presence of controls. In particular, in Column 1, where there are no other explanatory variables, the coefficient of ZAd is 0.022 with a t-statistic of 9.8. In Column 2, where distance, firm size and total advertisement expenses of the firm industry (i.e., $LogSize$ and $LogInduAD$) are added as controls, the coefficient of ZAd becomes 0.035 with a t-statistic of 4.53. Column 3 augments the probit regression also with city demographics, such as population ($LogPop$), income per capita ($LogIncPerCap$), median home values ($LogMedHome$) and unemployment rate ($Unemp$). Consequently, the coefficient slightly decreases to 0.031 with a t-statistic of 3.12.

The economic effect of the instrument on the probability with which a firm sports sponsors a city (defined as the marginal effect of ZAd times a 1 standard deviation increase) is shown in Panel B of Table 1.5. As stated in Column 1, without any controls, the impact of ZAd on the probability of sports advertising is 55.5 bps. Given that the average probability that a firm sports sponsors a given city is 107.8 bps, the effect corresponds to 51.6% of the mean. In Column 2, which controls for firm characteristics, the economic effect equals 35 bps or 32.5% of the mean. In Column 3, where city demographics are also added, the economic effect becomes equal to 21.5 bps or 20.0% of the mean. Therefore, the estimation results from the first-stage verify that conjecture that the instrument is a strong predictor of a firm's sports marketing

activities in a city.

Second Stage

Table 1.6 depicts the results from the Tobit regressions that correct for *SportAd* endogeneity using the control function method. In the odd numbered columns (i.e. Columns 1, 3, 5 and 7), the approximation of the correction functions Ψ_0 and Ψ_1 is quadratic, while in the even numbered columns (i.e Columns 2, 4, 6 and 8) the approximation is cubic. For a given set of controls, both orders of polynomial approximation yield similar results. Panel B, as before, depicts the estimated economic effect of sports advertisement and local bias on the household portfolio choice.

In Columns 1 and 2 of Panel A, the portfolio weights regressions on sports sponsorship control only for the local bias and the sports sponsorship endogeneity. In comparison to Column 1 of Table 1.4 (which does not correct for the sports sponsorship endogeneity), the estimated coefficient of *SportsAd* is dropped from 0.682 to 0.345 or 0.316, depending on the polynomial approximation order of the correction functions. The economic effect is substantially decreased from 32 bps to 9.35 bps or 6.48 bps. In Column 3 and 4, household demographics are added as controls. The coefficient of sports sponsorship again is decreased from 0.680 (in the uncorrected case) to 0.346 or 0.317. The implied economic effect goes down as before, from 32 bps goes to 9.31 bps or 6.41 bps.

Columns 5 and 6, further add stock characteristics as regressors. The coefficient of *SportsAd* is now reduced to 0.249 or 0.238 from 0.294. This time, the decrease in the economic effect of sports sponsorship is more moderate, from 4.79 bps to 3.7 bps or 3.42 bps. Column 7 and 8 incorporate also city demographics in the household investment equation. The estimated coefficient of *SportsAd* then becomes 0.262 or 2.55, with a respective economic effect of 3.98 bps or 3.73 bps. Based on Column 4 of Table 1.4, without including correction functions, the coefficient of sports sponsorship

is estimated to be higher (i.e., 0.300), as is the corresponding economic effect (i.e., 4.92 bps).

Based on the above results, the instrumentation of sports sponsorship reduces its impact on the portfolio, but not much. The fact that the new estimates of the *SportsAd* are smaller than before indicates indeed the possible presence of some firm catering to local tastes. However, in Column 7 and 8 of Table 1.6, where all observable fundamentals are included as controls in the portfolio weights regression, the reduction of the effect is not that sizable. In particular, the economic impact of sports sponsorship becomes equal to 49.9% to 46.8% of the average portfolio weight on a stock in the investment universe, from 61.7% in the uncorrected case.

Furthermore, in terms of the estimation of the coefficient of sports sponsorship, the instrumentation produces comparable results, even in the absence of some observable fundamentals as controls. That is, roughly, the estimates of *SportsAd* seem to be close with each other on average (across the first line in Panel A of Table 1.6). The coefficient of sports sponsorship also remains highly statistically significant. As for the coefficients of the controls, these are quite similar as before, with the local bias increasing very slightly.

1.5. Assessing the Advertising Over-Investment

1.5.1. Portfolio Trades

The censored regression framework for the household portfolio weights can also be used for the study of the sports sponsorship effect on the household trading activity. In the spirit of Barber and Odean (2000), I calculate the values of stock purchases and sales for every household in every month. I then scale the values of the households' monthly trades with their portfolio values in the respective month. In more detail, in a given month, household i's value of buys of stock j is given by the number of stock

j's shares that household i purchases (during that month) times the end-month price of stock j, divided by the household i's portfolio value at the end of the month, i.e.:

$$buy_{i,j,t} \equiv \frac{P_{j,t}\left(n_{i,j,t} - n_{i,j,t-1}\right)}{\sum\limits_{j \in \mathcal{J}} P_{j,t} n_{i,j,t}} \tag{1.6}$$

Analogously, in a given month, household i's value of sells of stock j is given by the number of stock j's shares sold by household i (during that month) times the price of the stock at the beginning of the month, over household i's portfolio value at the month's start date, i.e.:

$$sell_{i,j,t} \equiv \frac{P_{j,t-1}\left(n_{i,j,t-1} - n_{i,j,t}\right)}{\sum\limits_{j \in \mathcal{J}} P_{j,t-1} n_{i,j,t-1}} \tag{1.7}$$

In addition, the first and last monthly observations of a household's time series are dropped in the calculations. As a result, the value of a household's trades (relative to the portfolio value) is always bounded by zero and one.

As in the Tobit portfolio weight regressions, I run a censored regression for household i's buys and a censored regression for household i's sells in every month separately. I then again compute the average coefficient estimates and t-statistics (based on standard errors clustered at the household and city level). In the Tobit model of the buys, I let household i choose which (and how much) stock to purchase from all the available ones in the investment universe. On the other hand, since households do not short, in the Tobit model of the sells, I let household i choose which (and how much) stock to sell from the ones it holds at the end of the past month.

The estimation results from the household trading Tobit regressions are presented in Table 1.7. Columns 1 and 2 refer to household i's stock purchases. In Column 1, which does not correct for the sports sponsorship endogeneity, the coefficient of $SportsAd$ is estimated to be 0.091, with a t-statistic of 4.63. The implied economic effect is 0.10 bps or 47.2% of the average household stock buys (which is 0.21 bps).

In Column 2, in which the sports sponsorship endogeneity is corrected, the coefficient of *SportsAd* decreases to 0.084 with a t-statistic of 3.20. The implied economic effect consequently drops to 0.06 bps (or 28.3% of the average). Local bias is also estimated to have an effect on the households' stock purchases. With or without the control functions, the coefficient of *LogDist* equals roughly -0.05 and the implied economic effect varies from -0.02 bps to -0.03 bps (i.e., from -9.4% to -14.2% of the mean).

Columns 3 and 4 refer to household i's stock sales. In contrast to the estimation results in the household stock purchases, the coefficients of sports sponsorship and local bias are not found to be statistically significant. In particular, without the control functions, the estimated coefficient of *SportsAd* is -0.026, with a t-statistic of -0.84. When the control functions are incorporated, it becomes -0.053, with a t-stat of -0.44. In both cases, the negative sign indicates that the household is less likely to sell a stock that sports sponsors its city, but the t-statistics are low. The economic effect is also very moderate, ranging from -8.4% to -12.1% of the mean. Column 3 also shows that, without correcting for *SportsAd* endogeneity, the estimated coefficient of *LogDist* is 0.006 with a t-statistic of 0.98. In Column 4, which accounts for *SportsAd* endogeneity, the estimated coefficient of *LogDist* is 0.004 with a t-statistic of 0.33. In both cases, the economic effect of local bias is very small, varying from 2.1% to 1.3% of the mean.

The above figures show that sports sponsorship significantly affects a household's monthly buys. They also suggest, though, that the investment biases for stocks which sports sponsor the city of residence or are based geographically close matter more for the stock purchase decision than for the decision to sell. This finding is more consistent with theories pointing to the investors' limited attention. That is, if sports sponsorship or local bias were conferring information about a stock's profitable growth prospects, then a household would have been much less inclined to sell stocks with these features.

1.5.2. Portfolio Returns

I also examine whether the household investments in the sports sponsors outperform other Russell 1000 stock holdings. The methodlogoy that I follow is analogous to the one provided by Coval and Moskowitz (2001) and Cohen, Frazzini, and Malloy (2008). In every month, I classify each household's stockholdings into two distinct (household-specific) portfolios. The first portfolio is the "Sports Sponsorship Portfolio" and consists of holdings on stocks with sports advertisement in the city where the household resides. The second portfolio is the "Other Russell 1000 Portfolio" and is comprised by holdings on stocks which do not sponsor a sports teams in the household's MSA. I rescale the portfolio weights of every household in the two aforementioned portfolios, so that, for a given household, the weights in each portfolio sum to one. Households investing in only one of the two portfolios are also kept in the analysis.

Subsequently, for every household, I calculate the returns of the two portfolios in the next month. I then average the returns of each of the two portfolios across households, using their total portfolio values as weights, in order to obtain the monthly returns of the two portfolios.[36] Having obtained the household value-weighted returns of the "Sports Sponsorship Portfolio" and the "Other Russell 1000 Portfolio", I calculate the monthly payoffs from a strategy that longs the former and shorts the latter. The results of this household holdings performance test are shown in Table 1.8. Panel A presents the average monthly raw return of the strategy, while Panel B shows the corresponding Carhart 4-factor adjusted alpha.[37] In both panels, Newey-West standard errors with a lag order of 3 months are shown in brackets.

As reported in Column 1 in Panel A, the mean raw return from longing the "Sports Sponsorship Portfolio" and shorting the "Other Russell 1000 Portfolio" during

[36] Averaging household returns equally produces very similar results.

[37] The Carhart four factors are obtained from Kenneth French's website at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

the whole sample period is about 0.05%. The alpha is also very small (i.e. about 0.08%), as shown in Column 1 of Panel B. Neither the average return or the alpha are statistically significant. In Columns 2 and 3, I divide the sample into two equal sub-periods. Subsample 1 covers the period October 1992 - September 1994 and has a mean raw return of about -0.31% and an alpha of about -0.41%. On the other hand, Subsample 2 covers the period October 1994 - September 1996 with a mean raw return of approximately 0.40% and an alpha of about 0.49%. As before, in both subsamples, there is no statistical significance for the average return and the regression constant.

To better assess the above household performance figures, I examine the returns of the same strategy without conditioning on household investment or the city-specific advertisement exposure. That is, in every month, I construct the unconditional "Sports Sponsorship Portfolio", comprised by all firms that sports sponsor (without taking into account where they advertise). I compare the value-weighted return of that portfolio with the value-weighted return of the "Other Russell 1000 Portfolio".[38] The results of this unconditional performance test are depicted in Table 1.9.

As stated in Column 1, the mean raw return is about -0.09% with an alpha of approximately -0.04% for the whole sample period. Column 2 shows that in Subsample 1 the average return is about -0.27% and the regression constant is -0.22%. Column 3 shows that in Subsample 2 the return is on average 0.08%, while the alpha equals about 0.08% as well. There is no statistical significance in any of these figures.

For the whole sample period, the value-weighted strategy that conditions on household investment seems to do better than the unconditional one (i.e. about 0.05% vs. -0.09%). In Subsample 1, the mean raw return is worse (i.e. roughly -0.31% vs. -0.27%). Yet, in Subsample 2, the average return is much higher (i.e. 0.40% vs. 0.08%), as is the alpha (i.e. 0.49% vs. 0.08%).

In a nutshell, even though the strategies that condition on household investment

[38] Again, very similar results are also obtained for equally-weighted portfolios.

perform better than the unconditional ones, there is insufficient statistical evidence that firms with sports advertisement yield higher returns for the portfolio. A way to interpret this finding is that product market advertising is mainly relevant for capturing household attention and breeding familiarity than for conferring information about the prospects of a stock.

1.5.3. Directly vs. Non-Directly Recognizable Stocks

In the construction process of the sports sponsorship indicator variable, *SportsAd*, I assign the value 1 only to stocks which can be directly recognized from their product advertisement. To assess the degree to which the sports sponsorship effect on the household portfolio choice relies on this requirement, I introduce a new firm-city-specific indicator variable, namely *SportsAdDnD*, which equals 1 if the respective stock sports sponsors the corresponding city *regardless* of the firm's recognizability from product advertisement.[39] In that way, *SportsAdDnD* being equal to 1 is a necessary but not sufficient condition for *SportsAd* to equal 1 as well.[40]

I then replace the *SportsAd* with *SportsAdDnD* in Equations (1.1) and (1.5) and repeat the estimation. The results are depicted in Columns 1 and 2 of Table 1.11. As expected, since now a part of the sports sponsorship cannot be recognized by investors, the portfolio choice effect becomes weaker. More specifically, in Column 1, where the portfolio choice estimation does not account for the sports sponsorship endogeneity, the estimated coefficient of *SportsAdDnD* is 0.203 with a t-statistic of 5. The implied economic effect on the portfolio choice is 2.87 bps (or 36% of the mean). These figures are lower than the ones in Column 4 of Table 1.4. That is, with the *SportsAd* variable, the respective coefficient estimate is 0.300 and the economic effect 4.92 bps (or 61.7% of the mean).

[39] The "DnD" in *SportsAdDnD* stands for "direct or not direct" recognizability.

[40] The distribution of the non-directly recognizable sports sponsors over time is depicted in Table 1.10. For these stocks, $SportsAdDnD = 1$ but $SportsAd = 0$.

In Column 2, where the sports sponsorship endogeneity is being taken into account with control function polynomials, the estimated coefficient of *SportsAdDnD* equals 0.152 with a t-statistic of 2.87 and an implied economic effect of 1.95 bps (or 24.5% of the mean). Both the coefficient estimate and the economic effect are also smaller than the ones in Column 7 of Table 1.6. There, with the use of *SportsAd* as the sports sponsorship variable, the coefficient is 0.262 and the implied economic effect 3.98 bps (or 50% of the mean).

For a better assessment of a firm's direct recognizability requirement, I also estimate the portfolio choice Equations (1.1) and (1.5) by including both sports sponsorship variables, i.e. both *SportsAdDnD* and *SportsAd*. That is, I now perceive *SportsAdDnD* to be another firm-city observable fundamental, which can be introduced as an additional control in the portfolio weights regression on *SportsAd*. In that way, the new estimated coefficient of *SportsAd* can be interpreted as a difference effect.

Columns 3 and 4 of Table 1.11 depict the new estimation results. Without controlling for *SportsAd* endogeneity via the control function method, the coefficient estimate equals 0.405, which is higher than its previous value (i.e., 0.300). The new t-statistic is 7.05 and the new economic effect is 6.72 bps (or 84.32%), which is also higher than the effect without the incorporation of *SportsAdDnD* as a control (i.e., 4.92 bps). Column 4, in turn, takes into account the endogeneity of *SportsAd*. The coefficient estimate equals 0.374 and the economic effect is 5.88 bps (or 73.8% of the mean). Again both the estimated coefficient and economic effect are larger (i.e., relative to 0.262 and 3.98 bps from the previous estimation.)

Consequently, the direct recognizability of a firm from the product advertisements is important for the sports sponsorship to have a sizable effect on the household portfolio choice. This finding is also consistent with product advertising having a main attention or familiarity effect on the household investments.

1.6. Unbundling Familiarity

So far, the paper has identified the households' over-investment in stocks that sports sponsor their cities and presented evidence that suggests that the advertising over-investment is more consistent with theories of attention and familiarity. In this section, I re-estimate the household portfolio choice with the objective of showing how product market advertising interacts with local bias.

In the spirit of Table 1.1, I classify stocks into four classes of assets based on the geographical proximity of their headquarters from a household's residence and the sports sponsorship in a household's city. That is, from a household's perspective, stocks can be distant without sports advertisement (i.e., $DistNAd$), local without sports advertisement (i.e., $LocNAd$), distant with sports advertisement (i.e., $DistAd$) or local with sports advertisement (i.e., $LocAd$). I then replace the individual variables of sports sponsorship and local bias in Equations (1.1) and (1.5), namely $SportsAd$ and $LogDist$, with the aforementioned dummies (keeping $DistNAd$ as the base group).

1.6.1. Portfolio Demand

Table 1.12 depicts the new estimation results of the household portfolio choice. In Columns 1 and 2, stocks are characterized as local if their headquarters' ZIP Code is at most 250 miles away from a household's residential ZIP Code. Similarly, in Columns 3 and 4, stock locality is based on a 100-mile radius from the household's residential ZIP Code.

Across all columns, households are estimated to have the strongest preferences for stocks that are both local and sports sponsors of their residence city. Specifically, in Columns 1 and 3, where the endogeneity of sports sponsorship is not taken into account, the coefficient of $LocAd$, is estimated to be 0.767 for the 250 miles locality threshold and 0.846 for the 100 miles locality threshold. The corresponding economic

effects are 24.04 bps (i.e., 301.6% of the mean) or 31.27 bps (i.e., about 392.4% of the mean). Furthermore, in Columns 2 and 4, which account for the sports sponsorship endogeneity via the control function method, the coefficient of *LocAd* is decreased to 0.705 for the 250 miles locality threshold and 0.790 for the 100 miles locality threshold. The respective economic effects become 21.45 bps (or 269.1% of the mean) or 26.62 bps (or 334.0% of the mean).

In all the above cases, the estimated coefficient of *LocAd* is about twice as much higher than the coefficient of $LocNAd$, while the economic effect of the former is approximately 20 bps (i.e., 250% of the mean) higher than the economic effect of the latter.[41] That is, the familiarity from product market advertising drastically increases the the local bias that household exhibits.

As regards to the evaluation of the impact of sports advertising that non-local firms make, the coefficient of $DistAd$ across all columns is roughly 0.3, with an implied economic effect of roughly 40%, when taking into account the sports sponsorship endogeneity. That effect is big and shows that, by sports sponsoring a given city, firms with distant headquarters are more likely to be included (with a higher share) in that city's household portfolios.

In fact, it is possible that the sports sponsorship portfolio effect of distant firm can overpower the local bias for firms which do not sponsor a local team in the city where the household lives. Whether it actually does it or not depends on the distance threshold based on which a stock's locality is defined. In particular, for the threshold of 250 miles, distant stocks that sports sponsor a city are included with a higher probability and with a larger share in the portfolios of the city's households than local stocks that do not sponsor sponsor. In contrast, for the distance threshold of 100 miles, the situation is reversed.

[41] More precisely, accounting for the endogeneity of $SportsAs$ and considering stocks headquartered at most 250 miles away from a household's residential ZIP Code to be local, leads to a portfolio weight increase of $21.45 - 1.8 = 19.65$ bps. Analogously, for the 100-mile locality threshold, the portfolio weight increase equals $26.62 - 6.44 = 20.18$ bps.

In more detail, when a stock's locality is defined within a 250-mile radius from a household's residential ZIP Code and the sports sponsorship endogeneity is not corrected, the estimated coefficient of $DistAd$ is 0.340 with an implied economic effect 4.74 bps (or 59.5% of the average portfolio weight). The estimated coefficient of $LocNAd$, on the other hand, is estimated to be 0.268, with an implied economic effect of 2.77 bps (or 34.8% of the mean). The higher magnitude of $DistAd$'s impact vs. $LocNAd$'s impact is maintained, when I also account for $SportsAd$ endogeneity. In that case, the estimated coefficient of $DistAd$ drops to 0.296, while the coefficient of $LocNAd$ drops to 0.206. The economic effect for distant stocks that sports sponsor a household's city is 3.46 bps (or 43.4% of the mean), while, for local stocks which do not sports sponsor, it is 1.8 bps (or 22.6% of the mean).

For the distance threshold of 100 miles, the coefficient of $DistAd$ is 0.322 and 0.284, in the $SportsAd$ corrected and uncorrected case respectively. The corresponding economic effects are 4.63 bps (or 58.1% of the mean) and 3.17 bps (or 39.8% of the mean). Both coefficients and effects are now lower than the ones of $LocNAd$. The coefficient of $LocNAd$ equals 0.468 without the sports sponsorship endogeneity correction and 0.373 with the correction. The respective economic effects are 8.23 bps (or 103.3% of the mean) and 6.44 bps (or 80.8% of the mean).

1.6.2. Portfolio Diversification

Besides the effect on the household portfolio demand, I also explicitly examine the effect of product market advertising and local bias on the household portfolio (under-)diversification. In the spirit of Brandt, Santa-Clara, and Valkanov (2009), I replace the Tobit Equation (1.1) with the following linear regression:

$$w_{i,c,j} - w_j^{VW} = \alpha^{dev} + \beta_1^{dev} LocNAd_{c,j} + \beta_2^{dev} DistAd_{c,j} + \beta_3^{dev} LocAd_{c,j} + Controls_{i,c,j} + \tilde{\epsilon}_{i,c,j}^{dev} \tag{1.8}$$

where the dependent variable in the LHS of Equation (1.8) is the deviation of household i's portfolio weight on stock j from the value-weighted portfolio weight on that stock. The list of control variables is the same as before. The endogeneity of sports sponsorship in this case can also be corrected by following the same control function approach. In principle, under this new specification, the estimated coefficients can be directly interpreted as investment biases towards stocks with certain features. Equivalently, under the CAPM null, their coefficients should be equal to zero.

The estimated results of the household portfolio diversification are presented in Table 1.13. Overall, the same message is conferred as in Table 1.12. That is, regardless of the distance threshold for a stock's locality and the correction for the sports sponsorship endogeneity, households are expected to deviate from CAPM and overload their portfolios on stocks which sports sponsor their cities and are local.

More analytically, for the distance threshold of 250 miles and without accounting for the $SportsAd$ endogeneity, the coefficient of $DistAd$ equals 10.54 bps, thus being higher than the estimated coefficient of $LocNAd$, which is 9.39 bps. When the endogeneity of sports sponsorship is taken into account, the comparison of the two effects does not change. The coefficient of $DistAd$ decreases to 8.75 bps, but the coefficient of $LocNAd$ decreases to 6.53.

In contrast, for the distance threshold of 100 miles, the coefficient of $LocNAd$ is estimated to be higher than the one of $DistAd$. Without correcting for $SportsAd$ endogeneity, $LocNAd$ has an estimated effect of 17.14 bps, while $DistAd$ has an effect 12.45 bps. Once control functions are introduced in the regression, the effect of $LocNAd$ drops to 15.23 bps and the effect of $DistAd$ to 5.05 bps.

Moreover, for the distance threshold of 250 miles, the coefficient of *LocAd* equals 69.65 bps without the *SportsAd* endogeneity correction and 46.16 bps with the *SportsAd* endogeneity correction. For the distance threshold of 100 miles, the corresponding coefficients equal 84.96 bps and 62.81 bps respectively. Therefore, with the incorporation of the control functions, there is an increase in the household over-investment in local stocks ranging from $46.16 - 6.53 = 39.63$ bps to $62.81 - 15.23 = 47.58$ bps due to sports sponsorship.

1.7. Conclusion

Product market advertising increases a firm's visibility to consumers as well as to investors. However, the general unavailability of data on the exposure of investors to commercials poses challenges to the identification of the advertising effect on the portfolio choice. To overcome this issue, in this paper, I construct a dataset on the major league sports sponsorships of publicly listed firms, which I then merge with retail investment data from a large national discount broker.

I find that, even under the acknowledged presence of local equity preferences, a household's portfolio weight on a stock increases, if the stock sports sponsor the household's residence city. Sports sponsorship matters more for the household's stock purchase decisions than for the household's stock-selling decisions. Like local bias, it has a mute effect on the household's portfolio returns. Moreover, the effect of sports sponsorship on the portfolio choice is sizable, provided that the stock can be directly recognized from the product advertisement. Thus, theories that attribute the household over-investment in stocks that advertise to limited attention and familiarity are more consistent with data.

I also study how sports sponsorship interacts with local bias. I show that the household portfolio over-investment in local stocks can largely increase, if local stocks

sports sponsor a local team. But more importantly, with sports sponsorships, the portfolio share on distant firms increases substantially. The sports advertising effect for a distant firm can actually beat the local bias for a firm that is based within a 250-mile radius from the household's residence but does not sports sponsor.

In regard to future research, the extension of this study to accommodate the advertising of mutual funds (e.g., Barber, Odean, and Zheng (2005), Cronqvist (2006)) seems promising. Quite a few mutual funds (e.g., Fidelity, Putnam Investments, etc.) engage in sports marketing activities as they seek to increase their clientéle. Incorporating the city-fund sports sponsorship exposure in a household mutual fund demand framework is a feasible task.

Table 1.1: Sports Advertising and Local Bias in Household Stock Portfolios

This table summarizes the sports advertising over-investment and local bias in the stock holdings of households in the data. Panel A describes the local bias for stocks whose headquarters' ZIP Code lies within a 100-mile radius from a household's residential ZIP Code. Column 1A (Average Percentage of Sports Advertising in the Household Portfolio) reports the average share of a household stock-portfolio in firms that sports sponsor the household's residence city. It is computed as $\frac{1}{I}\sum_i \sum_j w_{i,c,j} SportsAd_{c,j}$, where $SportsAd_{c,j}$ is an indicator variable equal to 1 if firm j sponsors a sports team in city c where household i resides, $w_{i,j}$ is household i's portfolio weight on stock j, and I is the total number of households. Column 1B (Average Percentage of Sports Advertising in Excess of the Market) reports the difference between the average share of a household stock-portfolio in firms that sports sponsor the household's residence city and the average share of a value-weighted benchmark portfolio in the same stocks. It is computed as the difference between Column 1A and the average $\frac{1}{I}\sum_i \sum_j w_j^{VW} SportsAd_{c,j}$, where w_j^{VW} is a value-weighted benchmark portfolio weight on stock j based on the definition of the investment universe (and $SportsAd_{c,j}$ and I are as in Column 1A). The average t-statistic of the difference is also reported in the same column [in brackets]. Columns 2C and 2D are analogous to Columns 1A and 1B, but refer to firms which do *not* sports sponsor the household's residence city (replacing $SportsAd_{c,j}$ with $1 - SportsAd_{c,j}$ in the above formulas). Column 3 refers exclusively to local bias, summing up Column 1A with Column 2C, and Column 1B with Column 2D. In Panel's A Row 1, the summation formula is applied only to firms whose headquarters' ZIP Code is insides a 100-mile radius from the household's residential ZIP Code. In Panel's A Row 2, the summation formula is applied only to firms whose headquarters' ZIP Code is outside a 100-mile radius from the household's residential ZIP Code. Analogously, Panel B describes the local bias for stocks whose headquarters' ZIP Code lies within a 250-mile radius from a household's ZIP Code residence. The sample period is from September 1992 until August 1996. The universe of firms is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in Russell 1000 index during the sample period.

Panel A: Sports Advertising and Local Bias in 100 miles

	(1) SportsAd		(2) No SportsAd		(3) TOTAL	
Av. % of	Household Portfolio (A)	In Excess of Market (B)	Household Portfolio (C)	In Excess of Market (D)	Household Portfolio (A)+(C)	In Excess of Market (B)+(D)
$\leq$ 100 miles	5.3	+4.1 [21.6]	15.8	+10.6 [35.5]	21.1	+14.7 [43.5]
$>$ 100 miles	6.4	+1.3 [7.1]	72.5	-16 [-44.6]	78.9	-14.7 [-43.5]
TOTAL	11.7	+5.4 [21.3]	88.3	-5.4 [-21.3]	100	0

Panel B: Sports Advertising and Local Bias in 250 miles

	(1) SportsAd		(2) No SportsAd		(3) TOTAL	
Av. % of	Household Portfolio (A)	In Excess of Market (B)	Household Portfolio (C)	In Excess of Market (D)	Household Portfolio (A)+(C)	In Excess of Market (B)+(D)
$\leq$ 250 miles	6.1	+4.5 [22.3]	22	+11.4 [33.9]	28.1	+15.9 [43]
$>$ 250 miles	5.6	+0.9 [5.4]	66.3	-16.8 [-44]	71.9	-15.9 [-43]
TOTAL	11.7	+5.4 [21.3]	88.3	-5.4 [-21.3]	100	0

Table 1.2: Distribution of the Directly-Recognizable Publicly Traded Primary Sports Sponsors over Time

This table describes the distribution of the directly-recognizable publicly traded Primary Sports Sponsors over time. In every sports season, the publicly traded sports sponsors are identified from merging the Primary Sports Sponsors listed in the Sports Sponsor FactBook of Team Marketing Report (issued in 1993, 1994, 1995 and 1996) with the history files of CRSP. Directly-recognizable sponsors are sponsors which displayed their company name or logo in their product advertisement. (For a description of the distribution of the Non-Directly-Recognizable Publicly Traded Primary Sports Sponsors see Table 1.10.) Row 1 refers to their total number. Row 2 refers to their membership in Russell 1000. Row 3 refers to whether their securities are American depository receipts (ADR's). Rows 4-18 refer to their industrial classification based on their assignment in one of the 17 Fama-French industry portfolios. (The number of the respective industry portfolio is shown in parenthesis.) Food stands for food. Mines stands for mining and minerals. Oil stands for oil and petroleum products. Clths stands for textiles, apparel and footware. Durbl stands for consumer durables. Chems stands for chemicals. Cnsum stands for drugs, soap, perfumes and tobacco. Cnstr stands for construction and construction materials. FabPr stands for fabricated products. Machn stands for machinery and business equipment. Cars stands for automobiles. Trans stands for transportation. Rtail stands for retail stores. Finan stands for banks, insurance companies, and other financials. OthSvc stands for other and services. Due to lack of participation, the industry portfolios of Steel (9, i.e steel works etc.) and Utils (14, i.e. utilities) are not being displayed. Row 19 refers to the average number of Metropolitan Statistical Areas (MSA's) in which a publicly traded sponsor advertises and Row 20 refers to the median.

	Sports Season			
	1992-1993	1993-1994	1994-1995	1995-1996
Num. of Firms	167	166	175	174
Russell 1000	116	118	120	115
ADRs	18	15	17	16
Food (1)	17	19	17	14
Mines (2)	1	1	2	1
Oil (3)	17	15	14	14
Clths (4)	5	5	7	7
Durbl (5)	11	10	9	8
Chems (6)	1	0	2	2
Cnsum (7)	8	4	3	3
Cnstr (8)	4	5	4	3
FabPr (10)	1	1	0	0
Machn (11)	10	10	10	11
Cars (12)	10	8	8	8
Trans (13)	15	16	17	18
Rtail (15)	26	30	27	33
Finan (16)	18	18	31	29
OthSvc (17)	23	24	24	23
Av. MSAs	3.42	3.89	4.28	4.11
Med. MSAs	2	2	3	3

Table 1.3: Summary Statistics

This table reports the summary statistics of all the variables in the sample. Panel A refers to the household stock holdings. Portval is the portfolio value of a household. Numstk is the number of stocks that a household holds. Portwt is the portfolio weight of a household on a stock at the extensive margin (i.e. without conditioning on the household's positive positions). Panel B refers to the stock geographical proximity and sports advertisement in the investment universe. Dist is the distance between a household's address ZIP code and the address ZIP code of a stock's headquarters. LogDist is the log of Dist. Away250m (Away100m) is an indicator variable equal to 1 if a stock is headquartered more than 250 miles (100 miles) away from a household's residence. SportsAd is an indicator variable equal to 1 if a stock sponsors a sports team in the city of a household's residence and is directly recognizable by displaying the company name or logo in the product advertisement. Panel C shows the demographics of the households in the sample. Income is the income level. LogIncome is the log of Income. Age is the age of the household head. LogAge is the log of Age. Male is an indicator variable equal to 1 to if the head of a household is a male. FamSize is the number of members of a household. LogFamSize is the log of FamSize. Educ is the percentage of the household's ZIP Code population that holds a bachelor's or higher degree. White, Black, Hispanic, AsianOther are respectively the percentages of the household's ZIP Code population that are white, black, Hispanic and Asian or of other race. ProfProxim is the professional industrial proximity to a stock of the household's ZIP Code population. Panel D shows the demographics of the 71 Metropolitan Statistical Areas (MSAs) that were in the market for a sports team. in which households reside. Sportsteam is an indicator variable equal to 1 if the MSA has a team in the four major league sports, Pop is the population number, LogPop is the log of population, IncPerCap is the income per capita, LogIncPerCap is the log of IncPerCap, MedHome is the median home value, LogMedHome is the log of MedHome,Unemp is the unemployment rate. Panel E refers to the total advertising expenditure in the stocks' industries, as these are defined by the 17 Fama-French industry portfolios. Column 1 shows the average frequency of the industries in the sample. Column 2 shows the time-series average of their total advertising expenses (InduAD). Column 3 ranks the total advertising expenses of the industries in decreasing order (from 1 to 17). Panel F refers to the advertising expenditure of the stocks and Panel G shows their financial characteristics. For every financial characteristic, the summary statistics of the Sports Sponsors and the other firms in Russell 1000 are also depicted separately. Price is the price of a stock. Size is the market capitalization. LogSize is the log of Size. Book/Market is the book-to-market ratio. Turnover is the share turnover ratio. Momentum is the past 12-month return. Volatility is the volatility of the monthly returns in the past 12 months. Profitability is the ratio of past annual gross profits to assets. Investment is the past annual growth rate of assets. Sales is the past annual sales. LogSales is the log of Sales. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. The sample period is from September 1992 until August 1996.

Panel A: Household Stock Holdings					
	Mean	S.D.	Median	Min	Max
Portval ($ thous.)	38.44	143.36	13.30	1.00	11,990.15
Numstk	2.73	2.78	1.92	1	64
Portwt	7.97 bps	234.58 bps	0	0	1
Panel B: Stock Geographical Proximity and Sports Advertisement					
	Mean	S.D.	Median	Min	Max
Dist (stat. miles)	920.54	604.01	804.55	0	2,074.67
LogDist	6.43	1.16	6.69	0	7.64
Away250m	0.84	0.36	1	0	1
Away100m	0.93	0.26	1	0	1
SportsAd	0.02	0.14	0	0	1

Table Cont'd: Summary Statistics

Panel C: Household Demographics

	Mean	S.D.	Median	Min	Max
Income ($ thous.)	93.43	64.67	87.50	10.00	250.00
LogIncome	11.23	0.68	11.38	9.21	12.43
Age	54.04	14.00	50	21	80
LogAge	3.96	0.26	3.91	3.04	4.38
Male	0.89	0.32	1	0	1
FamSize	2.47	1.24	2	1	6
LogFamSize	0.78	0.5	0.69	0	1.79
Educ	0.36	0.15	0.36	0	0.91
White	0.81	0.18	0.86	0.01	1
Black	0.06	0.11	0.02	0	0.98
Hispanic	0.07	0.09	0.04	0	0.94
AsianOther	0.06	0.08	0.04	0	0.62
ProfProxim	0.08	0.06	0.07	0	0.76

Panel D: MSA Demographics

	Mean	S.D.	Median	Min	Max
Sportsteam	0.53	0.50	1	0	1
Pop (mil.)	2.07	2.66	1.20	0.26	17.70
LogPop	14.08	0.92	14.00	12.46	16.69
IncPerCap ($ thous.)	23.80	3.02	23.32	17.98	34.77
LogIncPerCap	3.16	0.12	3.15	2.89	3.55
MedHome ($ thous.)	97.25	30.36	89.84	50.60	215.11
LogMedHome	11.44	0.29	11.41	10.83	12.28
UnempRate (%)	5.17	1.43	5.00	2.28	10.35

Panel E: Industry Advertisement Expenditure

	Av. Frequency (%)	Av. InduAD ($ mil.)	Av. Ranking (1 to 16)
Food (1)	4.07	14,627.64	2
Mines (2)	1.54	40.28	16
Oil (3)	4.69	288.65	13
Clths (4)	1.99	1,816.33	10
Durbl (5)	2.66	7,200.11	6
Chems (6)	2.71	1,099.18	12
Cnsum (7)	4.33	14,042.03	3
Cnstr (8)	2.3	1,965.53	9
Steel (9)	1.63	145.43	14
FabPr (10)	0.92	52.56	15
Machn (11)	10.12	5,490.65	7
Cars (12)	1.95	13,193.14	4
Trans (13)	3.93	1,536.9	11
Utils (14)	7.81	0.04	17
Rtail (15)	7.95	10,644.09	5
Finan (16)	17.32	3,396.64	8
OthSvc (17)	24.08	15,441.5	1

Table Cont'd: Summary Statistics

Panel F: Firm Advertisement Expenditure						
		Mean	S.D.	Median	Min	Max
AD ($ mil.)	All	219.27	435.92	59.36	0.07	3,724
	Sports Sponsors	374.77	559.08	158.98	0.19	3,724
	Others	149.75	345.84	35.00	0.07	3,358
Panel G: Financial Characteristics of Stocks						
		Mean	S.D.	Median	Min	Max
Price ($)	All	34.79	24.37	30.13	0.05	739.75
	Sports Sponsors	40.62	26.93	36.25	0.05	261.25
	Others	33.89	23.83	29.63	0.19	739.75
Size ($ mil.)	All	3,522.47	7,676.05	1,293.31	0.70	144,224.20
	Sports Sponsors	9,832.16	15,726.51	3,628.3	0.70	144,224.20
	Others	2,548.10	4,768.07	1,216.68	0.79	86,346.17
LogSize	All	7.26	1.31	7.16	-0.36	11.88
	Sports Sponsors	7.77	2.14	8.2	-0.36	11.88
	Others	7.19	1.11	7.1	-0.23	11.37
Book/Market	All	6.07	113.41	0.47	-65.58	5,161.78
	Sports Sponsors	28.34	244.4	0.51	-6.84	4,040.82
	Others	2.63	74.4	0.46	-65.58	5,161.78
Turnover	All	0.10	0.13	0.06	0	4.55
	Sports Sponsors	0.09	0.13	0.06	0	4.55
	Others	0.10	0.14	0.05	0	2.95
Momentum	All	0.10	0.60	0.05	-0.93	46.16
	Sports Sponsors	0.07	0.36	0.05	-0.82	4.78
	Others	0.11	0.62	0.05	-0.93	46.16
Volatility	All	0.08	0.05	0.07	0	1.07
	Sports Sponsors	0.08	0.04	0.07	0	0.58
	Others	0.08	0.05	0.07	0	1.07
Profitability	All	0.33	0.27	0.27	-1.39	2.11
	Sports Sponsors	0.35	0.27	0.28	-0.03	1.39
	Others	0.33	0.27	0.27	-1.39	2.11
Investment	All	0.19	0.58	0.08	-0.84	31.29
	Sports Sponsors	0.13	0.51	0.06	-0.64	10.95
	Others	0.20	0.59	0.08	-0.84	31.29
Sales ($ mil.)	All	3,955.30	9,726.78	1,236.99	0	16,5370.20
	Sports Sponsors	13,732.73	21,312.17	6,514.07	0	165,370.20
	Others	2,445.43	4,695.12	1,070.59	0	78,069.20
LogSales	All	7.12	1.56	7.12	0	12.02
	Sports Sponsors	8.35	2.02	8.78	0	12.02
	Others	6.93	1.38	6.98	0	11.27

Table 1.4: Tobit Estimation of Household Portfolio Choice, No Correction for SportsAd Endogeneity

This table presents the estimation results from four Tobit models of household portfolio choice without correcting for the endogeneity of Sports Advertising (SportsAd). The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variables are SportsAd, i.e. an indicator variable equal to 1 if stock j sponsors a sports team in city c (where household i resides) and LogDist, i.e. the log of the distance between household i's ZIP Code and the ZIP Code of stock j's headquarters. Other variables include controls for household demographics (added in Column 2), stock characteristics (added in Column 3), and MSA demographics (added in Column 4). The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects of SportsAd and LogDist on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

Panel A: Coefficient Estimates in Portfolio Weight				
	(1)	(2)	(3)	(4)
SportsAd	0.682	0.680	0.294	0.300
	[14.81]	[15.02]	[7.06]	[7.69]
LogDist	-0.122	-0.121	-0.120	-0.127
	[-5.72]	[-5.81]	[-5.60]	[-6.27]
LogIncome		0.028	0.027	0.028
		[4.92]	[4.71]	[5.14]
LogAge		0.184	0.188	0.189
		[10.88]	[11.59]	[12.13]
LogFamSize		-0.007	-0.007	-0.006
		[-0.72]	[-0.83]	[-0.60]
Male		0.036	0.037	0.036
		[3.31]	[3.63]	[3.61]
Educ		0.124	0.112	0.160
		[3.33]	[3.29]	[5.09]
Black		0.023	0.018	0.044
		[0.53]	[0.39]	[1.19]
Hispanic		0.076	0.076	0.105
		[1.18]	[1.16]	[1.56]
AsianOther		0.066	0.033	0.034
		[0.93]	[0.45]	[0.59]
ProfProxim		0.418	0.283	0.212
		[2.28]	[3.16]	[2.88]
LogSize			0.122	0.121
			[6.80]	[6.59]

Table 1.4 Cont'd: Tobit Estimation of Household Portfolio Choice, No Correction for SportsAd Endogeneity

Panel A Cont'd: Coefficient Estimates in Portfolio Weight				
	(1)	(2)	(3)	(4)
Book/Market			0.000	0.000
			[0.69]	[0.57]
Turnover			0.407	0.398
			[12.20]	[11.58]
Momentum			-0.130	-0.130
			[-10.35]	[-10.38]
Volatility			3.075	3.066
			[25.66]	[25.66]
Profitability			0.017	0.016
			[0.80]	[0.76]
Investment			-0.041	-0.041
			[-4.37]	[-4.37]
LogSales			0.024	0.025
			[2.68]	[2.82]
LogSize × LogInduAD			0.021	0.021
			[12.20]	[11.76]
LogPop				-0.047
				[-3.78]
LogIncPerCap				-0.323
				[-2.71]
LogMedHome				0.087
				[2.05]
UnempRate				0.003
				[0.49]
Sportsteam				0.063
				[3.50]
Industry FE	NO	NO	YES	YES
Av. Households	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256

Panel B: Average Marginal Effects on Portfolio Weight in Basis Points and as Percentage of the Mean				
	(1)	(2)	(3)	(4)
SportsAd	32.46	32	4.79	4.92
	407.3%	401.5%	60.1%	61.7%
LogDist	-2.31	-2.27	-1.26	-1.33
	-29.0%	-28.5%	-15.8%	-16.7%

Table 1.5: First-Stage Probit Estimation of Firm Sports Advertising

This table presents the estimation results from three first-stage probit models of firm sports advertising. The estimation is performed in a panel of firms and cities for every year separately. Panel A presents the averages of the annual coefficient estimates and their t-statistics [shown in brackets] based on standard errors clustered at the firm and city level. The dependent variable is SportsAd, a dummy equal to 1 if firm j sponsors a sports team in city c. The key explanatory variable is Sportsteam×(LogSize×LogInduAD), i.e. the interaction of an indicator variable equal to 1 if city c has a sports team (Sporteam) in MLB, NBA, NFL or NHL and the interaction of the log of the market capitalization (LogSize) of firm j with the log of the total advertisement expenses in the industry (LogInduAD) of firm j. Firm industries are defined based on the assignment of their stocks in the 17 Fama-French industry portfolios. Column 2 adds as controls the log of average distance (LogDist) between the ZIP Code of firm j's headquarters and the ZIP Codes of city c, firm j's log of market capitalization (LogSize) and the log of the total advertisement expenses in the industry (LogInduAD) of firm j. Column 3 adds as additional controls the MSA demographics of city c. These include: the log of the population number (LogPop), the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). Panel B shows the economic effect of Sportsteam×(LogSize×LogInduAD) on SportsAd in basis points and as percentage of the mean, defined as the corresponding marginal effect times a 1 standard deviation increase. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. The set of cities consists of 71 MSAs which were in the market for a sports team.

Panel A: Coefficient Estimates of SportsAd			
	(1)	(2)	(3)
Sportsteam × $\underbrace{\text{LogSize} \times \text{LogInduAD}}_{\text{Advertising Propensity}}$	0.022	0.035	0.031
	[9.8]	[4.53]	[3.12]
LogDist		-0.228	-0.219
		[-8.17]	[-8.63]
LogSize		-0.092	-0.058
		[-1.04]	[-0.42]
LogInduAD		-0.15	-0.132
		[-2.72]	[-1.71]
LogPOP			0.300
			[6.05]
LogIncPerCap			-0.447
			[-1.21]
LogMedHome			0.251
			[1.87]
UnempRate			-0.055
			[-1.91]
Constant	-3.441	-1.738	-7.58
	[-22.67]	[-3.32]	[-4.52]
Av. Stocks	1,256	1,256	1,256
Av. Cities	71	71	71

Panel B: Econ. Effect on SportsAd in Basis Points and as Percentage of the Mean			
	(1)	(2)	(3)
Sportsteam × $\underbrace{\text{LogSize} \times \text{LogInduAD}}_{\text{Advertising Propensity}}$	55.5	35	21.5
	51.6%	32.5%	20.0%

Table 1.6: Tobit Estimation of Household Portfolio Choice, With Correction for SportsAd Endogeneity

This table presents the estimation results from eight Tobit models of household portfolio choice correcting for the endogeneity of Sports Advertising (SportsAd) using a control function method. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variables are SportsAd, i.e. an indicator variable equal to 1 if stock j sponsors a sports team in city c (where household i resides) and LogDist, i.e. the log of the distance between household i's ZIP Code and the ZIP Code of stock j's headquarters. In the odd-numbered columns (i.e. Columns 1, 3, 5 and 7), the correction functions Ψ_0 and Ψ_1 are approximated quadratically. In the even-numbered columns (i.e. Columns 2, 4, 6 and 8), the correction functions Ψ_0 and Ψ_1 are approximated cubicly. Other variables include controls for household demographics (added in Columns 3,4), stock characteristics (added in Columns 5,6), and MSA demographics (added in Columns 7,8). The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the coefficients in the correction functions Ψ_0 and Ψ_1, the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects of SportsAd and LogDist on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

Panel A: Coefficient Estimates of Portfolio Weight								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SportsAd	0.345	0.316	0.346	0.317	0.249	0.238	0.262	0.255
	[5.3]	[4.70]	[5.3]	[4.53]	[3.95]	[3.97]	[4.08]	[3.82]
LogDist	-0.127	-0.127	-0.124	-0.129	-0.138	-0.138	-0.131	-0.134
	[-5.75]	[-5.87]	[-5.75]	[-5.92]	[-8.16]	[-8.27]	[-7.19]	[-7.5]
LogIncome			0.022	0.022	0.028	0.028	0.028	0.028
			[3.55]	[3.54]	[4.92]	[4.91]	[5.14]	[5.14]
LogAge			0.164	0.164	0.192	0.193	0.189	0.189
			[9.22]	[9.25]	[12.15]	[12.22]	[12.21]	[12.21]
LogFamSize			-0.01	-0.01	-0.007	-0.007	-0.005	-0.005
			[-0.97]	[-0.97]	[-0.74]	[-0.76]	[-0.58]	[-0.59]
Male			0.046	0.046	0.035	0.035	0.035	0.035
			[3.52]	[3.54]	[3.52]	[3.5]	[3.53]	[3.54]
Educ			0.037	0.037	0.12	0.123	0.154	0.154
			[0.65]	[0.65]	[3.66]	[3.72]	[4.77]	[4.79]
Black			0.028	0.028	0.012	0.011	0.04	0.04
			[0.49]	[0.49]	[0.23]	[0.22]	[1.03]	[1.03]
Hispanic			-0.111	-0.112	0.11	0.116	0.103	0.104
			[-0.92]	[-0.92]	[1.51]	[1.61]	[1.54]	[1.58]
AsianOther			0.042	0.04	0.035	0.041	0.026	0.027
			[0.27]	[0.26]	[0.5]	[0.54]	[0.41]	[0.43]
ProfProxim			0.453	0.45	0.188	0.18	0.184	0.181
			[1.66]	[1.69]	[3.26]	[3.13]	[2.74]	[2.67]
LogSize					0.108	0.108	0.112	0.111
					[6.99]	[7.36]	[7.45]	[7.65]

Table 1.6 Cont'd: Tobit Estimation of Household Portfolio Choice, With Correction for SportsAd Endogeneity

Panel A: Coefficient Estimates of Portfolio Weight								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Book/Market					0.000	0.000	0.000	0.000
					[0.91]	[0.68]	[0.72]	[0.54]
Turnover					0.396	0.396	0.393	0.392
					[11.57]	[11.76]	[10.99]	[11.24]
Momentum					-0.133	-0.133	-0.129	-0.129
					[-10.54]	[-10.67]	[-10.1]	[-10.25]
Volatility					3.107	3.098	3.093	3.087
					[25.39]	[25.18]	[24.92]	[24.75]
Profitability					0.017	0.019	0.016	0.018
					[0.79]	[0.88]	[0.75]	[0.84]
Investment					-0.041	-0.041	-0.041	-0.041
					[-4.48]	[-4.51]	[-4.45]	[-4.48]
LogSales					0.026	0.026	0.026	0.026
					[3.16]	[3.19]	[3.13]	[3.18]
LogSize × LogInduAD					0.024	0.024	0.022	0.023
					[19.55]	[18.57]	[15.45]	[16.24]
LogPop							-0.044	-0.039
							[-2.57]	[-2.37]
LogIncPerCap							-0.299	-0.307
							[-2.65]	[-2.71]
LogMedHome							0.08	0.085
							[2.02]	[2.12]
Unemp							0.006	0.005
							[0.75]	[0.66]
Sportsteam							0.041	0.057
							[1.91]	[2.21]
Industry FE	NO	NO	NO	NO	YES	YES	YES	YES
Ψ^{1st}	YES	YES	YES	YES	YES	YES	YES	YES
Ψ^{2nd}	YES	YES	YES	YES	YES	YES	YES	YES
Ψ^{3rd}	NO	YES	NO	YES	NO	YES	NO	YES
Av. Households	9,462	9,462	9,462	9,462	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256	1,256	1,256	1,256	1,256

Panel B: Average Marginal Effects on Portfolio Weight in Basis Points and as Percentage of the Mean								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SportsAd	9.35	6.48	9.31	6.41	3.7	3.42	3.98	3.73
	117.3%	81.3%	116.8%	80.4%	46.4%	42.9%	49.9%	46.8%
LogDist	-2.37	-2.35	-2.31	-2.39	-1.41	-1.42	-1.35	-1.37
	-29.7%	-29.5%	-28.9%	-29.9%	-17.7%	-17.8%	-16.9%	-17.2%

Table 1.7: Tobit Estimation of Household Portfolio Buys and Sells, Without and With Correction for SportsAd Endogeneity

This table presents the estimation results from four Tobit models of household portfolio trades. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. Columns 1 and 2 refer to household i's buys in stock j relative to the portfolio value and the estimation is performed for all stocks in the investment universe. Columns 3 and 4 refer to household i's sells in stock j relative to the portfolio value and the estimation is performed for the stocks that household i holds at the end of the previous month. The key explanatory variables are SportsAd, i.e. an indicator variable equal to 1 if stock j sponsors a sports team in city c (where household i resides) and LogDist, i.e. the log of the distance between household i's ZIP Code and the ZIP Code of stock j's headquarters. Columns 1 and 3 do not correct for SportsAd endogeneity. Columns 2 and 4 correct for SportsAd endogeneity via the control function method. The correction functions Ψ_0 and Ψ_1 are approximated quadratically. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the coefficients in the correction functions Ψ_0 and Ψ_1, the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects of SportsAd and LogDist on $buy_{i,c,j}$ and $sell_{i,c,j}$ in basis points and as percentage of the mean. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

	Panel A: Coefficient Estimates			
	Buys		Sells	
	(1)	(2)	(3)	(4)
SportsAd	0.091	0.084	-0.026	-0.053
	[4.63]	[3.20]	[-0.84]	[-0.44]
LogDist	-0.053	-0.056	0.006	0.004
	[-5.38]	[-5.91]	[0.98]	[0.33]
LogIncome	0.016	0.016	-0.013	-0.013
	[2.03]	[2.02]	[-0.65]	[-0.65]
LogAge	0.060	0.059	-0.031	-0.031
	[2.87]	[2.86]	[-0.62]	[-0.61]
LogFamSize	0.001	0.001	0.000	-0.000
	[0.05]	[0.07]	[0.03]	[0.02]
Male	0.039	0.039	0.057	0.057
	[2.25]	[2.22]	[1.37]	[1.37]
Educ	0.078	0.076	0.032	0.029
	[1.90]	[1.84]	[0.43]	[0.38]
Black	-0.007	-0.008	-0.076	-0.077
	[-0.02]	[-0.03]	[-0.63]	[-0.64]
Hispanic	0.04	0.037	0.016	0.013
	[0.55]	[0.51]	[0.09]	[0.06]
AsianOther	0.088	0.081	0.221	0.216
	[1.27]	[1.17]	[1.81]	[1.76]
ProfProxim	0.125	0.110	0.115	0.096
	[1.55]	[1.38]	[0.82]	[0.69]
LogSize	0.077	0.069	0.008	0.001
	[4.00]	[3.66]	[0.19]	[0.06]

Table 1.7 Cont'd: Tobit Estimation of Household Portfolio Choice, With Correction for SportsAd Endogeneity

	Panel A: Coefficient Estimates Cont'd			
	Buys		Sells	
	(1)	(2)	(3)	(4)
Book/Market	-0.009	-0.008	-0.006	-0.006
	[-0.87]	[-0.75]	[-0.51]	[-0.46]
Turnover	0.304	0.293	0.254	0.251
	[10.19]	[9.47]	[2.99]	[2.95]
Momentum	-0.058	-0.054	0.023	0.021
	[-3.78]	[-3.5]	[1.1]	[1.00]
Volatility	1.110	1.094	0.748	0.757
	[9.04]	[9.1]	[2.72]	[2.77]
Profitability	0.035	0.034	0.080	0.079
	[2.05]	[2.06]	[1.66]	[1.64]
Investment	-0.001	0.000	0.015	0.015
	[0.29]	[0.44]	[0.98]	[0.98]
LogSales	-0.018	-0.015	-0.012	-0.011
	[-3.25]	[-2.82]	[-1.01]	[-0.95]
LogSize × LogInduAD	0.010	0.011	-0.002	-0.000
	[4.45]	[4.77]	[-0.42]	[-0.09]
LogPop	-0.015	-0.010	0.011	0.023
	[-1.39]	[-0.72]	[0.54]	[1.01]
LogIncPerCap	-0.126	-0.121	0.054	0.042
	[-1.29]	[-1.24]	[0.33]	[0.25]
LogMedHome	0.014	0.012	-0.042	-0.034
	[0.48]	[0.43]	[-0.76]	[-0.62]
UnempRate	0.002	0.003	-0.000	-0.002
	[0.39]	[0.49]	[-0.04]	[-0.17]
Sportsteam	0.022	0.011	-0.029	-0.022
	[0.95]	[0.47]	[-0.71]	[-0.5]
Industry FE	YES	YES	YES	YES
Ψ^{1st}	YES	YES	YES	YES
Ψ^{2nd}	YES	YES	YES	YES
Av. Households	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256

Panel B: Average Marginal Effects in Basis Points and as Percentage of the Mean				
	Buys		Sells	
	(1)	(2)	(3)	(4)
SportsAd	0.10	0.06	-6.82	-9.80
	47.2%	28.3%	-8.4%	-12.1%
LogDist	-0.02	-0.03	1.69	1.01
	-9.4%	-14.2%	2.1%	1.3%

Table 1.8: Household Returns of Primary Sports Sponsors vs. Other Russell 1000 Firms

This table presents the results from the returns test of publicly traded Primary Sports Sponsors vs. the other firms in Russell 1000 *conditional on household investment*. Returns are collected on a monthly basis and measured in percentage points (i.e. pps). *t*-statistics computed using Newey-West standard errors with lag order 3 are shown [in brackets]. The four Fama-French factors are used. Individual household returns are weighted by the respective total value of household holdings. Column 1 refers to the whole sample period, which is from October 1992 until September 1996. The month of October 1994 is used to break the sample into two equal subsamples. Columns 2 refers to Subsample 1. Column 3 refers to Subsample 2. Households reside in 71 MSAs which were in the market for a sports team.

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
	Panel A: Raw Returns		
Mean	0.046	-0.306	0.398
	[0.17]	[-0.91]	[1.03]
	Panel B: Return Regressions		
Mktrf	-0.040	0.032	-0.123
	[-0.54]	[0.24]	[-1.55]
SMB	-0.240	-0.025	-0.150
	[-3.49]	[-0.14]	[-1.54]
HML	0.327	0.458	0.473
	[4.50]	[2.66]	[2.91]
Mom	-0.064	-0.194	0.070
	[-0.53]	[-0.83]	[0.55]
Constant	0.082	-0.414	0.493
	[0.30]	[-1.38]	[1.04]

Table 1.9: Unconditional Returns of Primary Sports Sponsors vs. Other Russell 1000 Firms

This table presents the results from the unconditional returns test of the publicly traded Primary Sports Sponsors vs. the other stocks in Russell 1000. Returns are collected on a monthly basis and measured in percentage points (i.e. pps). *t*-statistics computed using Newey-West standard errors with lag order 3 are shown [in brackets]. The four Fama-French factors are used. Column 1 refers to the whole sample period, which is from October 1992 until September 1996. The month of October 1994 is used to break the sample into two equal subsamples. Column 2 refers to Subsample 1. Column 3 refers to Subsample 2.

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
	Panel A: Raw Returns		
Mean	-0.094	-0.267	0.079
	[-0.55]	[-1.05]	[0.34]
	Panel B: Returns Regressions		
Mktrf	-0.047	0.004	-0.091
	[-1.08]	[0.04]	[-1.36]
SMB	-0.259	-0.128	-0.227
	[-3.46]	[-0.94]	[-2.00]
HML	0.112	0.307	0.033
	[1.69]	[3.05]	[0.29]
Mom	-0.009	-0.170	0.082
	[-0.13]	[-1.49]	[0.74]
Constant	-0.035	-0.221	0.076
	[-0.22]	[-1.28]	[0.27]

Table 1.10: Description of the Non-Directly-Recognizable Publicly Traded Primary Sports Sponsors

This table describes the distribution of the non-directly recognizable publicly traded Primary Sports Sponsors over time. In every sports season, the publicly traded sports sponsors are identified from merging the Primary Sports Sponsors listed in the Sports Sponsor FactBook of Team Marketing Report (issued in 1993, 1994, 1995 and 1996) with the history files of CRSP. Non-directly recognizable sponsors are sponsors which did not displayed their company name or logo in their product advertisement. (For a description of the Directly-Recognizable Publicly Traded Primary Sports Sponsors see Table 1.2.) Row 1 refers to their total number. Row 2 refers to their membership in Russell 1000. Row 3 refers to whether their securities are American depository receipts (ADR's). Rows 4-14 refer to their industrial classification based on their assignment in one of the 17 Fama-French industry portfolios. (The number of the respective industry portfolio is shown in parenthesis.) Food stands for food. Clths stands for textiles, apparel and footware. Durbl stands for consumer durables. Chems stands for chemicals. Cnsum stands for drugs, soap, perfumes and tobacco. Machn stands for machinery and business equipment. Cars stands for automobiles. Trans stands for transportation. Finan stands for banks, insurance companies, and other financials. OthSvc stands for other and services. Due to lack of participation, the industry portfolios of Mines (2, i.e. mining and minerals), Cnstr (8, i.e. construction and construction materials), Steel (9, i.e. steel works etc.), FabPr (10, i.e. fabricated products), Rtail (13, i.e. retail stores) and Utils (14, i.e. utilities) are not being displayed. Row 15 refers to the average number of Metropolitan Statistical Areas (MSA's) in which a publicly traded sponsor advertises and Row 16 refers to the median.

	Sports Season			
	1992-1993	1993-1994	1994-1995	1995-1996
Num. of Firms	33	46	47	44
Russell 1000	22	29	30	26
ADRs	3	8	7	7
Food (1)	7	11	11	10
Oil (3)	1	2	2	2
Clths (4)	3	3	2	1
Durbl (5)	4	4	3	3
Chems (6)	1	3	3	1
Cnsum (7)	4	7	9	10
Machn (11)	1	1	1	1
Cars (12)	1	2	2	2
Rtail (15)	6	7	6	6
Finan (16)	0	1	1	2
OthSvc (17)	5	5	7	6
Av. MSAs	4.03	4.8	4.34	4.61
Med. MSAs	3	3	2	3

Table 1.11: Tobit Estimation of Household Portfolio Choice, Directly & Non-Directly Recognizable SportsAd

This table presents the estimation results from four Tobit models of household portfolio choice in the presence of directly and non-directly firm recognizable Sports Advertising (SportsAdDnD). The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variables are SportsAd, i.e. an indicator variable equal to 1 if stock j sponsors a sports team in city c (where household i resides) and is directly recognizable by displaying the company name or logo in the product advertisement, SportsAdDnD, i.e. an indicator variable equal to 1 if a stock sponsors a sports team in the city of a household's residence without conditioning on direct recognizability and LogDist, i.e. the log of the distance between household i's ZIP Code and the ZIP Code of stock j's headquarters. Columns 1 and 3 do not correct for the endogeneity of SportsAd and SportsAdDnD. Columns 2 and 4 correct for the endogeneity of SportsAd and SportsAdDnD using a control function method. The correction functions Ψ_0 and Ψ_1 are approximated quadratically. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the coefficients in the correction functions Ψ_0 and Ψ_1, the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects of SportsAd, SportsAdDnD and LogDist on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

Panel A: Coefficient Estimates in Portfolio Weight				
	(1)	(2)	(3)	(4)
SportsAd			0.405	0.374
			[7.05]	[6.31]
SportsAdDnD	0.203	0.152	-0.111	-0.161
	[5.00]	[2.87]	[-2.02]	[-2.45]
LogDist	-0.130	-0.134	-0.127	-0.131
	[-6.31]	[-7.20]	[-6.31]	[-7.18]
LogIncome	0.028	0.028	0.028	0.028
	[5.14]	[5.13]	[5.13]	[5.13]
LogAge	0.188	0.189	0.189	0.189
	[12.06]	[12.15]	[12.13]	[12.22]
LogFamSize	-0.006	-0.005	-0.006	-0.005
	[-0.6]	[-0.58]	[-0.6]	[-0.58]
Male	0.036	0.035	0.036	0.035
	[3.62]	[3.53]	[3.62]	[3.53]
Educ	0.160	0.154	0.160	0.154
	[5.01]	[4.70]	[5.07]	[4.77]
Black	0.043	0.039	0.044	0.040
	[1.15]	[1.01]	[1.18]	[1.03]
Hispanic	0.105	0.104	0.104	0.103
	[1.52]	[1.52]	[1.56]	[1.56]
AsianOther	0.043	0.031	0.038	0.026
	[0.67]	[0.47]	[0.60]	[0.41]

Table 1.11 Cont'd: Tobit Estimation of Household Portfolio Choice, Directly & Non-Directly Recognizable SportsAd

Panel A Cont'd: Coefficient Estimates in Portfolio Weight				
	(1)	(2)	(3)	(4)
ProfFam	0.209	0.185	0.208	0.184
	[2.83]	[2.78]	[2.83]	[2.75]
LogSize	0.125	0.118	0.118	0.111
	[6.81]	[7.65]	[6.53]	[7.31]
Book/Market	0.000	0.000	0.000	0.000
	[0.94]	[1.10]	[0.56]	[0.75]
Turnover	0.401	0.396	0.399	0.394
	[11.26]	[10.62]	[11.49]	[10.97]
Momentum	-0.130	-0.129	-0.130	-0.129
	[-10.36]	[-10.09]	[-10.38]	[-10.12]
Volatility	3.046	3.075	3.095	3.121
	[25.07]	[24.05]	[25.41]	[24.70]
Profitability	0.010	0.008	0.016	0.015
	[0.46]	[0.36]	[0.76]	[0.68]
Investment	-0.041	-0.041	-0.041	-0.041
	[-4.50]	[-4.56]	[-4.38]	[-4.44]
LogSales	0.024	0.025	0.027	0.028
	[2.59]	[2.78]	[3.10]	[3.34]
LogSize × LogInduAD	0.021	0.022	0.021	0.022
	[12.04]	[15.50]	[12.14]	[15.68]
LogPop	-0.048	-0.045	-0.048	-0.045
	[-3.66]	[-2.54]	[-3.66]	[-2.54]
LogIncPerCap	-0.335	-0.312	-0.335	-0.312
	[-2.76]	[-2.69]	[-2.76]	[-2.69]
LogMedHome	0.093	0.087	0.093	0.087
	[2.11]	[2.09]	[2.11]	[2.09]
UnempRate	0.003	0.005	0.003	0.005
	[0.45]	[0.70]	[0.45]	[0.70]
Sportsteam	0.066	0.046	0.066	0.046
	[3.59]	[2.10]	[3.59]	[2.10]
Industry FE	YES	YES	YES	YES
Ψ^{1st}	NO	YES	NO	YES
Ψ^{2nd}	NO	YES	NO	YES
Av. Households	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256

Panel B: Average Marginal Effects on Portfolio Weight in Basis Points and as Percentage of the Mean				
	(1)	(2)	(3)	(4)
SportsAd			6.72	5.88
			84.3%	73.8%
SportsAdDnD	2.87	1.95	-1.4	-1.91
	36.0%	24.5%	-17.6%	-24.0%
LogDist	-1.35	-1.37	-1.31	-1.33
	-16.9%	-17.2%	-16.4%	-16.7%

Table 1.12: Tobit Estimation of Household Portfolio with Local Bias and Sports Sponsorship Interactions

This table presents the estimation results from four Tobit models of household portfolio choice with local bias and sports sponsorship interactions. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variables are (i) LocNAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is less than or equal to 250 miles (100 miles) away from household i's address ZIP Code and stock j does not sponsor a sports team in household i's residence city, (ii) DistAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is more than 250 miles (100 miles) away from household i's address ZIP Code and stock j sponsors a sports team in household i's residence city, and (iii) LocAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is less than or equal to 250 miles (100 miles) away from household i's address ZIP Code and stock j sponsors a sports team in household i's residence city. Columns 1 and 2 define stock j's locality based on the threshold of 250 miles, while Columns 3 and 4 based on the threshold of 100 miles. Columns 1 and 3 do not correct for SportsAd endogeneity, while Columns 2 and 4 correct for the endogeneity of SportsAd using a control function method. The correction functions Ψ_0 and Ψ_1 are approximated quadratically. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the coefficients in the correction functions Ψ_0 and Ψ_1, the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects of LocNAd, DistAd and LocAd on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

	Panel A: Coefficient Estimates in Portfolio Weight			
	250-mile Radius		100-mile Radius	
	(1)	(2)	(3)	(4)
LocNAd	0.268	0.206	0.468	0.373
	[5.79]	[5.06]	[5.64]	[5.91]
DistAd	0.340	0.296	0.322	0.284
	[7.53]	[4.64]	[8.41]	[4.54]
LocAd	0.767	0.705	0.846	0.790
	[5.60]	[5.82]	[5.34]	[5.72]
LogIncome	0.027	0.027	0.027	0.027
	[4.99]	[4.98]	[5.01]	[5.00]
LogAge	0.188	0.188	0.188	0.188
	[12.09]	[12.12]	[12.10]	[12.17]
LogFamSize	-0.005	-0.005	-0.006	-0.006
	[-0.57]	[-0.60]	[-0.66]	[-0.64]
Male	0.034	0.034	0.037	0.035
	[3.48]	[3.44]	[3.73]	[3.58]
Educ	0.176	0.181	0.178	0.178
	[5.39]	[5.43]	[6.23]	[5.95]
Black	0.068	0.066	0.088	0.077
	[1.78]	[1.78]	[2.45]	[2.13]
Hispanic	0.103	0.112	0.079	0.094
	[2.04]	[2.22]	[1.65]	[2.00]
AsianOther	0.035	0.048	0.016	0.026
	[0.49]	[0.65]	[0.24]	[0.38]
ProfProx	0.338	0.344	0.306	0.302
	[3.26]	[3.41]	[3.41]	[3.34]

Table 1.12 Cont'd: Tobit Estimation of Household Portfolio with Local Bias and Sports Sponsorship Interactions

	Panel A: Coefficient Estimates in Portfolio Weight Cont'd			
	250-mile Radius		100-mile Radius	
	(1)	(2)	(3)	(4)
LogSize	0.125	0.134	0.118	0.122
	[8.92]	[10.56]	[6.56]	[9.08]
Book/Market	0.000	0.000	0.000	0.000
	[0.65]	[0.26]	[0.43]	[0.43]
Turnover	0.420	0.420	0.394	0.340
	[13.26]	[13.26]	[12.54]	[12.22]
Momentum	-0.130	-0.119	-0.131	-0.122
	[-9.85]	[-8.94]	[-9.96]	[-9.10]
Volatility	3.083	3.057	3.057	3.068
	[25.34]	[24.26]	[23.08]	[22.02]
Profitability	0.022	0.022	0.013	0.016
	[0.82]	[0.83]	[0.58]	[0.71]
Investment	-0.042	-0.040	-0.041	-0.041
	[-4.40]	[-4.23]	[-4.31]	[-4.36]
LogSales	0.023	0.022	0.025	0.024
	[2.85]	[2.75]	[2.93]	[2.85]
LogSize × LogInduAD	0.021	0.015	0.021	0.018
	[13.90]	[7.40]	[11.11]	[10.38]
LogPop	-0.031	-0.097	-0.030	-0.077
	[-2.68]	[-4.24]	[-2.34]	[-3.86]
LogIncPerCap	-0.210	-0.149	-0.182	-0.142
	[-2.04]	[-1.39]	[-1.66]	[-1.41]
LogMedHome	0.031	-0.007	-0.016	-0.031
	[0.81]	[-0.21]	[-0.53]	[-0.95]
UnempRate	-0.000	0.010	-0.004	0.006
	[0.02]	[1.19]	[-0.50]	[0.72]
Sportsteam	0.039	-0.023	0.029	-0.025
	[2.11]	[-1.01]	[1.62]	[-1.13]
Industry FE	YES	YES	YES	YES
Ψ^{1st}	NO	YES	NO	YES
Ψ^{2nd}	NO	YES	NO	YES
Av. Households	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256

Panel B: Average Marginal Effects on Portfolio Weight in Basis Points and as Percentage of the Mean				
	250-mile Radius		100-mile Radius	
	(1)	(2)	(3)	(4)
LocNAd	2.77	1.8	8.23	6.44
	34.8%	22.6%	103.3%	80.8%
DistAd	4.74	3.46	4.63	3.17
	59.5%	43.4%	58.1%	39.8%
LocAd	24.04	21.45	31.27	26.62
	301.6%	269.1%	392.4%	334.0%

Table 1.13: Regressions of Household Portfolio Weight Deviations from the Market with Local Bias and Sports Sponsorship Interactions

This table presents the estimation results from four linear regressions of household portfolio weight deviations from the market, with local bias and sports sponsorship interactions. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates in basis points (bps) and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j} - w_j^{VW}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variables are (i) LocNAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is less than or equal to 250 miles (100 miles) away from household i's address ZIP Code and stock j does not sponsor a sports team in household i's residence city, (ii) DistAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is more than 250 miles (100 miles) away from household i's address ZIP Code and stock j sponsors a sports team in household i's residence city, and (iii) LocAd, an indicator variable equal to one if stock j's headquarters' ZIP Code is less than or equal to 250 miles (100 miles) away from household i's address ZIP Code and stock j sponsors a sports team in household i's residence city. Columns 1 and 2 define stock j's locality based on the threshold of 250 miles, while Columns 3 and 4 based on the threshold of 100 miles. Columns 1 and 3 do not correct for SportsAd endogeneity, while Columns 2 and 4 correct for the endogeneity of SportsAd using a control function method. The correction functions Ψ_0 and Ψ_1 are approximated quadratically. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographic controls refer to household i and include: the log of the household's income (LogIncome), the log of the age (LogAge) of the household's head, the log of the family size (LogFamSize) of the household and an indicator variable equal to one if the household has a male head (Male). They also contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White) and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics refer to stock j and consist of: the log of market capitalization (LogSize), the book-to-market ratio (Book/Market), the monthly turnover ratio (Turnover), the past annual return (Momentum), the standard deviation of monthly returns in the past year (Volatility), the ratio of past annual gross profits to assets (Profitability), the past annual growth rate of assets (Investment), the log of past annual sales (LogSales) and the interaction of the log of market capitalization (LogSize) with the log of the total advertisement expenses in the industry of the stock (LogInduAD). They also include industry indicator variables based on the assignment of stock j in one of the 17 Fama-French industry portfolios. Controls for MSA demographics refer to city c (where household i resides) and include: the log of the population number (LogPop) of city c, the log of the income per capita (LogIncPerCap), the log of the median home price (LogMedHome) and the unemployment rate (UnempRate). The estimated values of the coefficients in the correction functions Ψ_0 and Ψ_1 and the constant in the linear regressions are not reported for brevity. The sample period is from September 1992 until August 1996. The universe of stocks is comprised by the publicly traded Primary Sports Sponsors and all the stocks that were ever included in the Russell 1000 index during the sample period. Households reside in 71 MSAs which were in the market for a sports team.

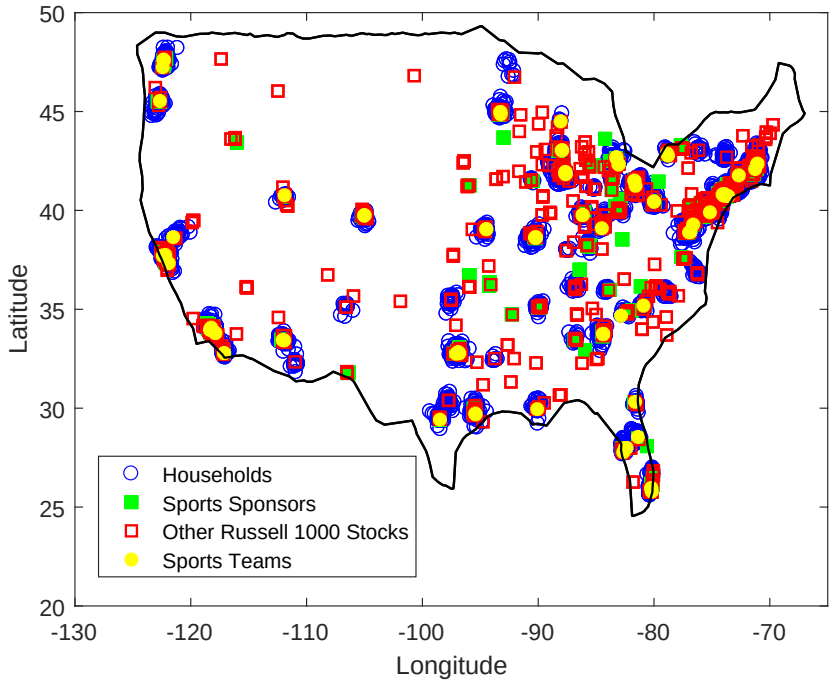
	Deviations from VW Market Port. Weights (bps)			
	250-mile Radius		100-mile Radius	
	(1)	(2)	(3)	(4)
LocNAd	9.39	6.53	17.14	15.23
	[4.55]	[4.20]	[3.62]	[3.69]
DistAd	10.54	8.75	12.45	5.05
	[2.56]	[2.10]	[3.36]	[2.26]
LocAd	69.65	46.16	84.96	62.81
	[2.46]	[2.72]	[2.2]	[2.49]
LogIncome	-0.01	0.00	-0.01	-0.01
	[-0.50]	[0.04]	[-0.85]	[-0.42]
LogAge	0.14	0.16	0.12	0.13
	[1.54]	[1.38]	[1.57]	[1.49]
LogFamSize	-0.04	-0.05	-0.05	-0.05
	[-1.11]	[-1.20]	[-1.73]	[-1.57]
Male	-0.03	-0.04	0.02	0.01
	[-1.08]	[-0.93]	[1.01]	[0.35]
Educ	1.10	1.36	1.09	1.22
	[2.89]	[2.82]	[3.92]	[3.56]
Black	0.32	0.27	0.74	0.60
	[0.94]	[0.78]	[1.35]	[1.17]
Hispanic	2.16	2.61	1.54	1.93
	[1.48]	[1.49]	[1.29]	[1.48]
AsianOther	1.37	2.11	0.96	1.45
	[1.34]	[1.54]	[1.12]	[1.33]

Table 1.13 Cont'd: Regressions of Household Portfolio Weight Deviations from the Market with Local Bias and Sports Sponsorship Interactions

	Deviations from VW Market Port. Weights (bps)			
	250-mile Radius		100-mile Radius	
	(1)	(2)	(3)	(4)
ProfProx	14.06	14.53	13.83	13.79
	[4.81]	[5.07]	[5.43]	[5.52]
LogSize	-5.40	-4.71	-5.46	-5.10
	[-14.40]	[-10.98]	[-11.51]	[-12.52]
Book/Market	0.00	0.00	0.00	0.00
	[-1.40]	[-1.04]	[-1.58]	[-0.88]
Turnover	11.39	11.85	10.61	10.96
	[8.21]	[8.66]	[8.92]	[9.26]
Momentum	-1.90	-1.67	-1.90	-1.78
	[-6.95]	[-6.64]	[-7.20]	[-7.13]
Volatility	43.63	41.89	42.6	41.62
	[9.64]	[8.52]	[9.08]	[8.34]
Profitability	2.15	2.27	1.93	2.05
	[2.14]	[2.22]	[2.05]	[2.11]
Investment	-1.00	-0.93	-0.99	-0.93
	[-5.60]	[-4.92]	[-5.62]	[-5.04]
LogSales	-0.65	-0.68	-0.67	-0.66
	[-3.52]	[-3.64]	[-3.23]	[-3.17]
LogSize × LogInduAD	0.57	0.34	0.58	0.44
	[12.13]	[4.10]	[10.77]	[6.91]
LogPop	-0.71	-2.49	-0.74	-1.78
	[-2.36]	[-3.21]	[-2.63]	[-2.90]
LogIncPerCap	-6.41	-5.41	-5.08	-4.63
	[-2.86]	[-2.08]	[-2.35]	[-1.94]
LogMedHome	1.38	0.75	0.20	-0.05
	[1.52]	[0.73]	[0.30]	[-0.03]
UnempRate	-0.12	0.11	-0.22	-0.05
	[-0.73]	[0.42]	[-1.55]	[-0.41]
Sportsteam	1.09	0.68	0.89	0.64
	[2.73]	[1.42]	[2.97]	[1.55]
Industry FE	YES	YES	YES	YES
Ψ^{1st}	NO	YES	NO	YES
Ψ^{2nd}	NO	YES	NO	YES
Av. Households	9,462	9,462	9,462	9,462
Av. Stocks	1,256	1,256	1,256	1,256

Figure 1.1: Geographical Distribution of Households, Stocks and Sports Teams

This figure depicts the geographical coordinates of 18,421 households, 286 publicly traded Primary Sports Sponsors, 1,123 other public firms which were ever included in the Russell 1000 index during that period and 107 sports teams. The address ZIP-codes of the households, the stocks' headquarters and the teams' headquarters are converted to geographical coordinates based on the geographical relationship files provided by the US Census Bureau. The horizontal axis shows the longitude coordinates and the vertical axis shows the latitude coordinates. Blue circles indicate the address ZIP codes of the households. Green squares indicate the address ZIP codes of the headquarters of the publicly traded Primary Sports Sponsors. Red squares indicate the address ZIP codes of the headquarters of the other Russell 1000 firms. Yellow circles indicate the address ZIP codes of the sports teams. The sample period is from September 1992 until August 1996. The list of the publicly traded Primary Sports Sponsors is extracted in the Sports Sponsor FactBook of Team Marketing Report (issued in 1993, 1994, 1995 and 1996) and the Russell 1000. Households reside in 71 MSAs which were in the market for a sports team.



Chapter 2

Location Choice, Portfolio Choice*

2.1. Introduction

A long-standing puzzle in financial economics is that households hold undiversified stock portfolios tilted toward firms headquartered near where they reside. Contrary to a value-weighted market portfolio prescription of the CAPM (Sharpe (1964)), households load on local stocks regardless of their market value. In canonical regressions of household stock-portfolio weights on demographic and stock characteristics, *distance* from household residence to firm headquarters emerges as a key explanatory variable. This local-bias appears in many countries.[1] This phenomenon is even more puzzling in some ways than its better-known cousin, the international home-bias puzzle, where households in different countries tilt toward stocks in their own country (French and

*Co-authored with Harrison Hong, John R. Eckel Jr. Professor of Financial Economics in the Department of Economics at Columbia University, and Jiangmin Xu, Assistant Professor of Finance at the Guanghua School of Management of Peking University.

We thank Ulrich Mueller, Mark Watson, Chris Sims, Bo Honore, Matti Keloharju, Selale Tuzel, Baolian Wang, Kirill Evdokimov, Motohiro Yogo, Atif Mian, Jakub Kastl, and participants at Econometrics and Finance seminars at Princeton University, University of Toronto, Johns Hopkins University, Columbia University, University of Maryland, University of Oxford, the 2016 LACEA/LAMES Conference, the 2016 China Five-Star Workshop in Finance, the 2016 NYU Shanghai Volatility Institute Conference, the 2017 WFA Conference, the 2017 CICF Conference, and the 2017 Helsinki Finance Summit for helpful comments.

[1]Prominent studies include the US (Zhu (2002)), Finland (Grinblatt and Keloharju (2001)) and China (Feng and Seasholes (2008)) to name a few.

Poterba (1991)). In the international setting, portfolio costs or restrictions at least seem plausible impediments toward diversification.

Given the potentially high costs of under-diversification for households, many theories have been given for this local bias in the literature.[2] The leading interpretation of local bias is a familiarity heuristic (e.g., Heath and Tversky (1991), French and Poterba (1991), Huberman (2001)), whereby investors favor local stocks that they view as having competence—be it the company they work for, companies near them that they know friends at, or even the telephone company that services their homes. The most recent studies using the best practice portfolio return methodologies find no evidence that local stock picks of households out-perform their distant stock picks, thereby pointing to the familiarity heuristic as the main causal mechanism as opposed to an alternative explanation having to do with rational information acquisition (see, e.g., Seasholes and Zhu (2010)).[3]

We point out in this paper that extant empirical work on local bias subtly but crucially assume that households locate randomly, which is likely to be counterfactual. Notably, in endogenous location choice models from the literature on urban and real estate economics, agents optimally locate in cities that provide them with the highest utility (e.g. Bajari and Kahn (2005) and Bayer, Ferreira, and McMillan (2007)). This sorting or self-selection depends on observable factors, such as the congruence of city characteristics and household demographics, as well as on factors that are unobserv-

[2]There is a sizeable literature examining the potential costs of under-diversification in stock portfolios and financial mistakes or literacy more generally (see, e.g., Campbell (2006), Bayer, Bernheim, and Scholz (2009), Agarwal, Driscoll, Gabaix, and Laibson (2009), Lusardi and Mitchell (2007)). Many households around the world have concentrated local stock holdings and little diversification through other investment vehicles (see, e.g., Keloharju, Knupfer, and Rantapuska (2012)). The costs of foregone diversification would seem to be large unless their local stock picks can significantly outperform the market.

[3]This stands in contrast to professional investors, where Coval and Moskowitz (2001) find such an informational advantage in the local trades of mutual funds managers. Another causal explanation for local bias of households is Keeping up with the Joneses' preferences (Luttmer (2005), Charles, Hurst, and Roussanov (2009)) which leads to demand for local stocks a form of hedging (DeMarzo, Kaniel, and Kremer (2004), Gomez, Priestley, and Zapatero (2009), Hong, Jiang, Wang, and Zhao (2014)).

able to the econometrician (i.e. latent factors), such as latent subjective household expectations about the economic prospects of a city. Households that are identical in every observable dimension locate in different cities if they hold heterogenous enough subjective expectations. Such optimal spatial sorting models are consistent with migration patterns that one sees in the US (e.g. Bishop (2007), Kennan and Walker (2011), Kaplan and Schulhofer-Wohl (2012), Diamond (2016)).

Even though empirical studies control for detailed observable outcomes such as occupational status of the household or professional proximity (e.g., Massa and Simonov (2006)), they cannot account for latent subjective expectations about the economic prospects of different cities. Households naturally prefer to move to areas which they view as having a bright economic future not only for themselves but for their family or future generations. But these latent expectations in their location choice are likely to be naturally correlated down the line with optimism about local versus distant stocks to the extent stocks are sensitive to economic conditions of the region of their firm headquarters. More importantly, their latent expectations about cities that they did not move to (as we demonstrate below) play an important role in their portfolio choices as well. In other words, to what extent does proximity play a causal role in local bias and to what extent does it simply reflect selection bias? As a thought experiment, if we were to randomly locate households in different cities, would they still exhibit the same degree of local bias? The location selection bias is ultimately an omitted-variables problem, whereby unobservable location factors correlated with investment demand shocks are ignored, violating the strict exogeneity assumption on distance in a standard portfolio weights regression.

As such, we develop a methodology to account for the effect of endogenous location decisions as a form of selection bias on household portfolio choice. We start by using a location choice model from urban economics in which households choose their residence based on a match between their own demographics (e.g., age and

family status) and the demographics of the location (e.g., urban vs. rural). This match congruence is captured by interaction terms involving city demographics and household demographics. Examples including households with kids wanting to live in less urban areas or areas with more affordable housing.

Our optimal location choice model allows us to recover the expected location utility of a household in a city and hence the probability that it locates there. Similar to Heckman (1977), these location probabilities can then be added in the portfolio weights regression as extra covariates that capture unobserved locational shocks. Our exclusion restriction is that the match congruence involving certain MSA demographics and household characteristics, which explain location choice for non-stock investment rationales, do not affect portfolio choice other than through our selection/optimal location choice mechanism. We elaborate in Section 4 on the plausibility of this exclusion restriction. To the extent that there is no location selection bias, introducing these probabilities should not affect the estimate of the coefficient on distance.

To quantify the effect of our endogenous location decision adjustment on the local bias of portfolio choices, we apply our model to a US brokerage database with roughly 11,000 households living in 57 MSAs with a population above 750K, during the period of 1991-1996. This is a widely used sample (Odean (1999), Barber and Odean (2000)) in which high income households have a significant fraction of their assets in stocks. As the investment universe, we consider stocks that belong to the Russell 1000 Index, an index that includes the largest 1000 stocks in the stock market based on market capitalization.[4]

We first run the standard censored regression of local bias: a Tobit regression of household stock portfolio weights on stock characteristics, household and city demo-

[4]The benchmark of Russell 1000 is purely for tractability, given the high degree of household portfolio sparsity. In the online Appendix, we show that our results apply also for the extended investment universe of Russell 3000.

graphics (e.g. income, occupation), and, notably, the distance between a household's residence and a stock's headquarters. We find a large distance effect, in terms of both economic and statistical significance, consistent with analyses in the literature.

To correct for a potential bias of location selection, we estimate an optimal location choice model for the same households using MSA demographics data. The results are similar to what is found in the literature on optimal location decisions. Namely, the interaction of MSA demographics with household characteristics plays a crucial role in explaining location decisions and preferences. This matching structure between the demographics of a household and the demographics of a city is consistent with the fact that households locate in a city mainly due to family, employment and housing-related reasons (e.g., 2016 U.S. Census Population Survey[5]) For instance, higher income, managerial, and white collar households are more likely to locate in high population centers. Older and blue collar households are more likely to locate in MSAs with high unemployment.

With the estimates of this locational choice model, we impute the probabilities with which each household locates in the various cities. We then plug these probabilities as additional controls in our portfolio weights regression, in line with the robust identification assumptions that we consider for our correction function. Across a wide variety of specifications, we find a large drop of around 50% in the effect of distance on the portfolio weight on a stock once we account for selection.

These reduced form estimates suggest a structural model of household portfolio choice and local bias depending on household priors about stocks in a city, which we can proxy using Pearson residuals from our location choice model, and familiarity, which we model as investor confidence about stock-payoff signals increasing with proximity to the stock headquarter. Households are risk neutral but face short-sales constraints and quadratic trading costs. These assumptions are realistic for our

[5]See https://census.gov/newsroom/press-releases/2016/cb16-189.html.

investors since they hold concentrated positions, do not short and face high trading costs.

Households form their portfolios based on both priors about stocks in a city and also i.i.d. private signals about individual stock payoffs. We model stock-payoff signals in the standard linear-normal signal extraction framework. The signal might end up being positive or negative, but we prove that the expected portfolio weight of a stock increases with prior and proximity. The reason is censoring due to short-sales constraints: as a household's private-signal precision increases, he places more weight on these signals relative to his prior. Since there are shorting constraints, we only see the positive latent demand from better priors and the increased precision and hence the higher expected portfolio weight. Thus, our model provides two rationales for local bias: optimistic priors and familiarity. We estimate our model and gauge the relative importance of these two factors. We find that the household priors account for 60% of local bias and familiarity or distance 40% using goodness-of-fit tests.

2.2. Accounting for Location Choice in Reduced Form Portfolio Choice Regressions

In this section, we present a simple framework that highlights the implications of a household's location choice on its subsequent investment decisions. We index households with i, stocks with j, and periods with t; overall, there are T periods in each of which live I_t households that can potentially invest in J stocks. The total number of cities, throughout the years, is C, and we denote with c the city in which household i resides, and with h the city in which stock j is headquartered.

2.2.1. Location Choice

Since in our data households do not move, we only model their location choice in the beginning of their time series. In line with a standard discrete choice model, we decompose the utility that household i derives from a city $\ell = 1, \ldots, C$ into the sum of an observable component, $V_{i,\ell}$, and an unobservable idiosyncratic shock, $e_{i,\ell}$, and assume that household i is a utility maximizer locating to a city c satisfying the following relationship:

$$c = \arg \max_{\ell \in \{1,\ldots,C\}} \{V_{i,\ell} + e_{i,\ell}\} \tag{2.1}$$

Household i's observable utility from city ℓ is a linear combination of the city's characteristics at the time at which the location decision is made, which we group into a $K \times 1$ vector $\boldsymbol{z}_\ell$. In our empirical analysis, this vector consists of the city's population, unemployment rate, income per capita and house price index. On the other hand, household i's unobservable utility from city ℓ refers to location factors that we, as econometricians, cannot observe, such as subjective expectations. Moreover, $e_{i,\ell}$ also refers to other factors during the location decision process that are uncorrelated with latent demand for stocks.

Although households in a given period view the same city characteristics, they interpret them differently, i.e.:

$$V_{i,\ell} = \boldsymbol{\rho}_i \boldsymbol{z}_\ell \tag{2.2}$$

where $\boldsymbol{\rho}_i$ is the vector of household i's responses. In particular, we assume observed heterogeneity in preferences through a matching structure. That is, we decompose $\boldsymbol{\rho}_i$ into a component that is common across all households, $\boldsymbol{\rho}$, and a component that linearly depends on household i's $M \times 1$ vector of demographics, $\boldsymbol{D}_i$, (through a $K \times M$ matrix of parameters $\boldsymbol{\Pi}$) i.e.:

$$\boldsymbol{\rho}_i = \boldsymbol{\rho} + \boldsymbol{\Pi} \boldsymbol{D}_i \tag{2.3}$$

The vector of household i's demographics, $\boldsymbol{D}_i$, that we use in our empirical analysis has as elements its total income, age, job code, gender, marital status and number of kids. By combining Equations (2.2) and (2.3), we eventually represent household i's observed utility from locating in city ℓ as:

$$V_{i,\ell} = \underbrace{\sum_{k=1}^{K} \rho_k z_{\ell,k}}_{\delta_\ell} + \underbrace{\sum_{k=1}^{K} \sum_{m=1}^{M} \pi_{k,m} D_{i,m} z_{\ell,k}}_{\mu_{i,\ell}} \tag{2.4}$$

where δ_ℓ is the observed utility from the characteristics of city ℓ that is common for all households, while $\mu_{i,\ell}$ is the observed utility from the characteristics of city ℓ which is different across households. Equation (2.4) implies that once we estimate the location parameters $\boldsymbol{\theta}^{loc} \equiv (\boldsymbol{\rho}, \boldsymbol{\Pi})$ from the data, we will have also estimated the observed utilities of household i from all the available locations, $\{V_{i,\ell}\}_{\ell=1,\ldots,C}$.

Next, we define household i's maximum order statistic with respect to a city c as:

$$v_{i,c} = \max_{\ell \in \{1,\ldots,C\}/c} \{V_{i,\ell} - V_{i,c} + e_{i,\ell} - e_{i,c}\} \tag{2.5}$$

so that household i's location rule in Equation (2.1) can be rewritten as:

$$r_{i,c} = \mathbf{1}\left[v_{i,c} < 0\right] \tag{2.6}$$

where $r_{i,c}$ denotes household i's decision to reside in city c and $\mathbf{1}[\cdot]$ is an indicator function. Assuming that, conditional on the observables, household i's idiosyncratic shocks, $\{e_{i,\ell,t}\}_{\ell=1}^{C}$, are independently and identically distributed according to the extreme value type I distribution, we can calculate the probability with which it resides in city c as follows:

$$p_{i,c} \equiv \mathbb{P}\left(v_{i,c} < 0 \left| \{V_{i,\ell}\}_{\ell=1}^{C} \right.\right) = \frac{\exp\left(V_{i,c}\right)}{\sum_{\ell=1}^{C} \exp\left(V_{i,\ell}\right)} \tag{2.7}$$

2.2.2. Reduced-Form Portfolio Choice Regressions

We consider a widely-used regression specification in the literature, which we will then micro-found in Section 6. Since we will estimate portfolio parameters for every period separately (thus allowing for time variation in households' portfolio preferences and expectations), we omit the period subscript t in the discussion that follows. Specifically, we assume that household i, residing in city c, decides how much to invest in stock j, headquartered in city h, according to a linear factor rule censored at zero:

$$w_{i,c,h,j} = \left(\alpha + \boldsymbol{\beta}\boldsymbol{x}_j + \boldsymbol{\gamma}\boldsymbol{D}_i + \delta dist_{i,c,h,j} + \epsilon_{i,c,h,j}\right)^+ \tag{2.8}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ captures both household i's extensive and intensive margin[6], $\boldsymbol{x}_j$ is the vector of stock j's financial characteristics - in particular, its price, size, book-to-market ratio, turnover, momentum, volatility and industry code, $\boldsymbol{D}_i$ is the vector of household i's demographics exactly as in its location choice problem, while $dist_{i,c,h,j}$ is the distance between household i's ZIP-code in city c and stock j's headquarters ZIP-code in city h. Moreover, $\epsilon_{i,c,h,j}$ is household i's idiosyncratic demand shock for stock j, when the former resides in city c and the latter is headquartered in city h. For instance, it could refer to whether household i thinks highly of stock j due to the stock's board of directors or due to the fact that the stock belongs to a specific industry sector that household i really likes (e.g. technology).

This empirical censored specification is appropriate for capturing the high degree of sparsity that household portfolios exhibit (including on average only 2 stocks in

[6] Households do not short.

Russell 1000). In line with a Tobit model, we assume that, conditional on all observables, these errors are distributed according to the normal distribution. In the spirit of Petersen (2009), when we estimate the model, we cluster standard errors at the household and city (aka MSA) level. Without correcting for location selection, the conditional mean of $\epsilon_{i,c,h,j}$ is assumed to be zero. errors are i.i.d. according to a conditional on the observables, these errors are i.i.d. according to a zero mean normal distribution with variance σ^2. Hence, the portfolio parameters to be estimated from the data are $\boldsymbol{\theta}^{port} \equiv (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta)$, with δ being the main parameter of interest (i.e. the coefficient on the distance variable).

2.2.3. Selection Correction

The distance between household i's ZIP-code in city c and the ZIP-code of stock j in city h where it is headquartered can always be expressed as a function of (i) the distance between household i's ZIP-code and the central ZIP-code of city c in which it resides, $dist_{i,c}$, (ii) the distance between the central ZIP-code of city c in which it resides and the central ZIP-code of city h in which stock j is headquartered, $dist_{c,h}$, and (iii) the distance between the central ZIP-code of city h in which stock j is headquartered and stock j's headquarters ZIP-code, $dist_{h,j}$. In short, denoting $S(\cdot)$ this function, we can write that:

$$dist_{i,c,h,j} = S\left(dist_{i,c}, dist_{c,h}, dist_{h,j}\right) \tag{2.9}$$

The need to correct for selection arises from the fact that the distance between the central ZIP-code of city c in which household i resides and the central ZIP-code of city h in which stock j is headquartered is the *outcome* of household i's location choice. That is, as long as household i is not randomly assigned to the city where it resides, the location rule in Equation (2.6) implies that:

$$dist_{c,h} = \sum_{\ell=1}^{C} dist_{\ell,h} r_{i,\ell} \tag{2.10}$$

where every distance between the central ZIP-code of a city ℓ and the central ZIP-code of city h in which stock j is headquartered, $dist_{\ell,h}$, is multiplied by household i's respective indicator function of its decision to live there, $r_{i,\ell} = \mathbf{1}[v_{i,\ell} < 0]$. Having that in mind, it is very likely that $\epsilon_{i,c,h,j}$, i.e. household i's idiosyncratic investment error when it lives in city c and considers investing in stock j headquartered in city h, is correlated with the idiosyncratic location errors, $\{e_{i,\ell}\}_{\ell=1}^{C}$ - especially $e_{i,c}$ and $e_{i,h}$ - as these are summarized by the maximum order statistic of the city c where household i actually resides, $v_{i,c}$. To show such a potential correlation more clearly, we decompose $\epsilon_{i,c,h,j}$ as follows:

$$\epsilon_{i,c,h,j} = \mathbb{E}\left(\epsilon_{i,c,h,j} \left| v_{i,c} < 0, \{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) + \eta_{i,c,h,j} \tag{2.11}$$

where $\eta_{i,c,h,j}$ is an idiosyncratic stock-city investment error which, by construction, is independent of household i's location decision to reside in city c. That is, $\eta_{i,c,h,j}$ is mean-zero given all observables. As for the conditional expectation of the original idiosyncratic investment error, $\epsilon_{i,c,h,j}$, given household i's decision to live in city c and the observed location utilities $\{V_{i,\ell}\}_{\ell=1}^{C}$, which are estimated from the location choice model in a first stage, it can be calculated as follows:

$$\begin{aligned} \mathbb{E}\left(\epsilon_{i,c,h,j} \left| v_{i,c} < 0, \{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{0} \frac{\epsilon_{i,c,h,j} f\left(\epsilon_{i,c,h,j}, v_{i,c} \left| \{V_{i,\ell}\}_{\ell=1}^{C}\right.\right)}{\mathbb{P}\left(v_{i,c} < 0 \left| \{V_{i,\ell}\}_{\ell=1}^{C}\right.\right)} dv_{i,c} d\epsilon_{i,j} \\ &= \psi_{c,h}\left(\{V_{i,\ell}\}_{\ell=1}^{C}\right) \end{aligned} \tag{2.12}$$

where $\psi_{c,h}(\cdot)$ is an *unknown* correction function whose actual form depends on assumptions regarding the *joint* distribution of $\epsilon_{i,c,h,j}$ and $v_{i,c}$. The value of the correc-

tion function is in principle non-zero, unless $\epsilon_{i,c,h,j}$ and $v_{i,c}$ are independent.[7] Consequently, based on Equations (2.9) to (2.12), the distance variable in the portfolio choice regression, $dist_{i,c,h,j}$, is correlated with the original investment idiosyncratic error, $\epsilon_{i,c,h,j}$, through the correction function $\psi_{c,h}(\cdot)$. Any estimation procedure that ignores this correlation is destined to yield biased estimates on the respective coefficient, δ.

To avoid a selection bias in the distance coefficient, one necessarily has to restrict the structure through which the observed utilities, $\{V_{i,\ell}\}_{\ell=1}^{C}$, enter the function $\psi_{c,h}$. We implement a robust non-parametric method following Dahl (2002).

To this end, we invoke the monotonic relationship between household i's observed location utilities, $\{V_{i,\ell}\}_{\ell=1}^{C}$, and its location probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, which allows us to write:

$$\psi_{c,h}\left(\{V_{i,\ell}\}_{\ell=1}^{C}\right) = \Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) \tag{2.13}$$

where now we have a new *unknown* correction function, namely $\Psi_{c,h}(\cdot)$, in terms of location probabilities. The probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, in the correction function capture the impact of unobservable location factors on subsequent investment decisions, *given* residence choice.

Hence, combining Equations (2.8), (2.11), (2.12) and (2.13) yields the portfolio choice regression equation that corrects for selection based on location:

[7] In that case, in the numerator of Equation (2.12), we have that:

$$\begin{aligned}\int_{-\infty}^{+\infty}\int_{-\infty}^{0} \epsilon_{i,c,h,j} f\left(\epsilon_{i,c,}v_{i,c}\left|\{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) dv_{i,c} d\epsilon_{i,c,h,} &= \int_{-\infty}^{+\infty} \epsilon_{i,c,h,j} f\left(\epsilon_{i,c,h,j}\left|\{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) d\epsilon_{i,c,h,j} \int_{-\infty}^{0} f\left(v_{i,c}\left|\{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) dv_{i,c} \\ &= \mathbb{E}\left(\epsilon_{i,c,h,j}\left|\{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) \mathbb{P}\left(v_{i,c}<0\left|\{V_{i,\ell}\}_{\ell=1}^{C}\right.\right) \\ &= 0\end{aligned}$$

since the conditional mean of $\epsilon_{i,c,h,j}$, given the observables, is zero.

$$w_{i,c,h,j} = \left(\alpha + \boldsymbol{\beta}\boldsymbol{x}_j + \boldsymbol{\gamma}\boldsymbol{D}_i + \delta dist_{i,c,h,j} + \Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) + \eta_{i,c,h,j}\right)^{+} \tag{2.14}$$

As in Dahl (2002), since more than one probabilities enter the correction function, the selection bias on the distance coefficient cannot be *ex ante* assessed. The censoring of the portfolio weights regression and the ability of a household to invest in both local and distant stocks further contribute to this.[8] Intuitively though, under the premise that latent location preferences for an area (here, captured by the correction function) are positively correlated with unobservable latent investment preferences for stocks headquartered there, we anticipate an over-estimation of the local bias when location selection is ignored.

2.3. Data

2.3.1. MSA Demographics

The data on the demographics of the Metropolitan Statistical Areas (MSAs) are gathered at a quarterly frequency from various sources. Specifically, information on the MSAs' population and unemployment rate is gathered from the Bureau of Labor Statistics (BLS), while information on their total income is collected from the Bureau of Economic Analysis (BEA). The house price index (HPI) is taken from the Federal Housing Finance Agency (FHFA). We focus on MSAs whose population at the end of 1996 was at least 750,000.[9] As Coval and Moskowitz (1999), we also drop Alaska, Hawaii and Puerto Rico. These filters leave us with 57 MSAs in total.

[8] In principle, the selection correction will also affect the coefficient estimates of household demographics due to the matching structure in location choice.

[9] This requirement is for tractability and to make sure that we have enough residents from the investment database to estimate their location probabilities more precisely. In the online Appendix, we obtain similar results extending our analysis to 80 MSAs with a population higher than 500K in the beginning of 1991.

The summary statistics of the MSA characteristics for all the quarters in the sample are presented in Panel A of Table 2.1. The mean population is 2.5 million with a standard deviation of 2.8 million. The mean unemployment rate is 5.8%. The mean income per capita is 23.6 thousand dollars, while the mean HPI is 99.2.

2.3.2. Household Demographics

Our household investment data are drawn from the database of a national discount brokerage firm. See Barber and Odean (2000) for detailed descriptions. The dataset is an unbalanced panel of month-end account statements from approximately 78,000 households at the stock level (CUSIP). The sample period spans from January 1991 to November 1996. Most households have multiple accounts which we aggregate, in order to obtain their total long positions in a given stock. As Ivković and Weisbenner (2005) report, the majority of accounts is non-retirement (e.g., cash or investment) and the few retirement accounts do not refer to 401(k) plans. Therefore, mechanical effects on the stock choice from the shares of an employer are limited.

In our analysis, we omit households whose demographic information is incomplete. Specifically, we require households to have a non-missing address ZIP code, income, family size, age, gender and marital status of the head. This criterion decreases the sample size to approximately 40,000. We also require the observability of the job code of a household's head, according to which its occupation is classified as (i) professional or technical, (ii) administrative or managerial, (iii) sales or service, (iv) white-collar or (v) blue-collar. That criterion further reduces the number of household to around 17,500.

Based on these data and our definitions for the location and investment universe below, we derive our final sample, which consists of 10,712 unique households with complete information on demographics and stock portfolios. These households do not move across MSAs, but stay in their original location either until the last date in the

data or until they close their accounts.[10] Their first time-series observations comprise the sample of our location choice model.

The summary statistics of the households' demographics are presented in Panel B of Table 2.1. The income of households in our sample has a mean of 82.4 thousand dollars and a median of 87.5 thousand dollars. This is to be expected since only households with sufficient income would participate in the stock market to begin with. The mean age is 47.7 years. About 55% of the households are professionals and 28% are managerial. Sales service, white collar and blue collar accounts comprise about 8%, 5% and 4% of the total respectively. Approximately, 93% of the households are headed by a male and 79% of the heads are married. On average, households have at most 1 kid.

2.3.3. Geographical Distribution of Household Stock Holdings

The universe of stocks that we examine is the Russell 1000 Index. We focus only on stocks located in the same 57 MSAs as above, with a complete list of financial characteristics as described below. This filter leads us to a total number of 1,056 different stocks for the whole period. Using the geographical coordinates of the ZIP-code of every household and the ZIP-code of the headquarters of every stock, we calculate their spherical distances, which are the key variable in our study.[11] [12]

[10] According to the US Census Bureau, the average percentage of movers during our sample period (1991-1996) was on average about 17% (http://www.census.gov/newsroom/press-releases/2015/cb15-47.html). This means that, roughly, a household would be expected to change residence every 6 ($\approx 1/0.17$) years. Given that our own sample period is six years, we expect that only few households in the data moved.

[11] We measure distance in degrees. Multiplying by $2\pi R/360$ converts it to miles (kilometers), where $R \approx$3,963 miles (6,378 kilometers).

[12] Since Compustat contains only the most recent headquarters' addresses of the stocks, a variety of sources (e.g., EDGAR, COMPHIST, etc.) is utilized to ensure that the headquarter information in the sample period is accurate. Pirinsky and Wang (2006) identify 118 firm relocations from 1992 to 1997 and Tuzel and Zhang (2017) about 300 from 1990 to 2005. In most cases, the firms that moved were small and not members of Russell 1000.

The geographical distribution of the households in our sample and the stocks in Russell 1000 is presented in Figure 2.1 via a map of latitude and longitude coordinates of the households and the stocks' headquarters. This figure depicts the geographical coordinates of the 10,712 households and the 1,056 stocks in Russell 1000. The address ZIP-codes of households and the stocks' headquarters are converted to geographical coordinates based on the correspondence provided by the US Census Bureau. The horizontal axis is in longitude coordinate, while the vertical axis is in latitude coordinate. The blue circles indicate households, while the red squares indicate stocks. From this map of the geographical distribution of households and stocks in our sample, we can see the high numbers of households and stocks located in the New York area (latitude around 40, longitude around -75) and California (latitude around 35, longitude around -120), which is what one would expect.

The portfolio positions of households are summarized in Panel C of Table 2.1. The mean value of a household's portfolio in common stocks is about $40,000 (averaged across time periods from 1991 to 1996), while the median value is about $15,000.[13] The standard deviation of stock holdings across our sample is $152,000. In terms of trading, as Barber and Odean (2000) document, these households are quite active on average, buying and selling 6-7% of their stock portfolio every month, and turning over 75% of it every year.

Furthermore, on average, a household in our sample has a portfolio weight of 11.2 bps on a Russell 1000 stock and holds 1.85 stocks. The standard deviation of the number of stocks is 1.55, indicating that most of the households in the sample are under-diversified, even as their stock holdings comprise a substantial fraction of their assets. The median number of stocks that a household holds is 1, while the standard deviation of a portfolio weight is 0.03. In the same panel, we also report the mean

[13]The report by the US Census Bureau on net worth and asset ownership of households in 1998 and 2000 shows that in 1998, the median value of holdings in stocks for a typical US household is $16,800 (https://www.census.gov/prod/2003pubs/p70-88.pdf). This information indicates that our sample is similar to the stock holding situation of US households in the '90s.

distance of a household's residence to a Russell 1000 firm headquarters, which is 17.6. The standard deviation is 11.45. These figures can be compared to the distance between California and New York, which is approximately 57 degrees. In addition, we also construct dummies indicating whether a stock is headquartered more than 250 miles and 100 miles away from a household's residence. The average percentage of households that are away from a stock's headquarters according to these metrics is about 90%.[14]

2.3.4. Stock Financial Characteristics

For each stock that belonged to the Russell 1000 Index during our sample period, we gather financial characteristics from CRSP and Compustat. In particular, we generate quarterly time series for their price, market value, book-to-market ratio, turnover, past twelve-month momentum and volatility. We construct these characteristics as Gompers and Metrick (2001). We also invoke the Fama-French industry classification of stocks into 17 categories based on the four-digit SIC code, which is available from Kenneth R. French's website.[15]

We summarize all stock characteristics in Panel D of Table 2.1. The mean price of a stock in the Russell 1000 index is about $34. The mean market capitalization is around 3.4 billion dollars. The mean book-to-market ratio is 0.43. The mean share turnover is 0.3. The mean past 12-month return is 0.17 and the mean monthly return volatility is 0.18. The industrial composition of the Russel 1000 Index is reflected by the 17 Fama-French industry classification; 26% of the stocks belong to the "Other" industry category, 16% of them are in "Finance" (referring to banks, insurance companies and other financials), while 12% belong to the "Machines" category (for machinery and business equipment).

[14] These thresholds are also used by Ivković and Weisbenner (2005) and Seasholes and Zhu (2010).
[15] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

2.3.5. Summary Statistics on Local Bias of Stock Portfolios

The summary statistics of the local bias (LB) in our household stock holdings data are given in Table 2.2 and are constructed as in Coval and Moskowitz (1999). Column 2 (labeled "Avg. Distance from Holdings") reports the average *portfolio weighted distance* of households from their stock holdings, defined as $\frac{1}{I}\sum_i \sum_j w_j^i d_j^i$, where d_j^i is the ZIP-code distance between a household i's residential area and the headquarters area of a stock j, w_j^i is the household i's portfolio weight on stock j, and I is the total number of households. Column 3 (labeled "Avg. Distance from Benchmark") reports the average portfolio weighted distance of households from the Russell 1000 benchmark portfolio, computed as $\frac{1}{I}\sum_i \sum_j \bar{w}_j d_j^i$, where $\bar{w}_j$ is the Russell 1000 benchmark portfolio weight on stock j. Row 1 has as benchmark the equally weighted portfolio, while Row 2 refers to the value-weighted portfolio. Column 5 (labeled "Difference") reports the average difference between Column 3 and Column 4, which is essentially the average local bias of households in distance units. Column 6 (labeled "% Bias (LB)") reports the local bias (LB) measure as a percentage. Column 7 reports the t-statistics for the LB measure. Independent of which benchmark is used (the values are about the same), the local bias is always high in terms of both magnitude and statistical significance. Specifically, using the equally weighted portfolio, the local bias is 8.29 or 45.45%, while, using the value-weighted portfolio, it is slightly decreased to 8.26 or 43.72%.[16]

[16]The percentage LB is more than four times the local bias that Coval and Moskowitz (1999) report for non-index fund managers in 1995.

2.4. Identification

2.4.1. Exclusion Restriction

As it is the case with any Heckman (1977) correction model (e.g. Puhani (2000)), to avoid identification through functional forms, an exclusion restriction is required, according to which at least one of the factors affecting location choice does not directly show up in the portfolio choice regression. Based on our location choice model, a household decides in which city to reside, taking into account the demographics of all the available locations as well as how it matches to them through its own demographics that are driven by non-stock investment reasons. The instruments in our portfolio choice regressions are derived through this matching process. Household decide where they want to live based for life-style choices, such as singles wanting to live in urban centers to meet other singles, or housing affordability issues such as families wanting to live in places with affordable housing. We use these rationales or predictors as instruments. That is, our identification assumption is that the way through which certain household's demographics interacting with certain demographics of a city has no direct impact on its stock investment decisions, other than through our mechanism. It does not matter of locational decisions can also be driven by other financial or human capital considerations (Ortalo-Magné and Prat (2010) and Hizmo (2015)). It is only important that we find variables which predict location choice that can be plausibly excluded from the portfolio choice.

2.4.2. Specification of the Correction Function

Based on Equation (2.14), we need to evaluate C^2 correction functions, $\Psi_{c,h}(\cdot)$, each of which has C probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, as arguments. This high dimensionality issue makes the estimation of the portfolio choice regression infeasible. As a first remedy,

in line with Dahl (2002), we adopt the following two identification assumptions:[17]

Assumption 1 (Two Index Sufficiency): The correction function has only two arguments, namely the probability with which household i resides in city c and the probability with which it resides in city h:

$$\Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) = \Psi_{c,h}\left(p_{i,c}, p_{i,h}\right) \tag{2.15}$$

Assumption 2 (Residence City Independence): The form of the correction function does not depend on the residence city c unless $h = c$, i.e.:

$$\Psi_{c,h}\left(p_{i,c}, p_{i,h}\right) = \left\{ \begin{array}{c} \Psi_{c}\left(p_{i,c}\right) \text{ if } h = c \\ \Psi_{h}\left(p_{i,c}, p_{i,h}\right) \text{ if } h \neq c \end{array} \right\} \tag{2.16}$$

According to Assumption 1, only two out of C probabilities are relevant for the location correction, namely the probability that household i locates in the area in which it actually resides, $p_{i,c}$, and the probability that household i locates in the city that has the headquarters of the stock in which it considers investing, $p_{i,h}$. However, that assumption still leaves us with C^2 correction functions. To this end, we impose Assumption 2, which reduces the total number of correction functions to its square root. In particular, it assumes that a correction function should not depend on the identity of the city in which household i resides. Of course, if it happens that stock j is located in the same city as household i does, i.e. $h = c$, then the identity of the residence city becomes relevant again.

Lastly, because the total number of cities in our data is large, i.e. $C = 58$, we are still left with a high number of correction functions to be estimated from the data.

[17] The difference in our context is that the location choice - portfolio choice model is more complex than the mobility-earnings model studied in Dahl (2002), since, in principle, a household can purchase stocks headquartered in any city, regardless of where it resides. On the other hand, in a mobility-earnings model, a household receives wage offers only from its own city. This extra element leads us to additionally impose Assumption 3, in order to make the estimation feasible, as we discuss in text.

To this end, we conveniently impose the following assumption:

Assumption 3 (Homogeneity): The form of the correction function is not stock headquartered city-specific, i.e.:

$$\begin{aligned} \Psi_c\left(p_{i,c}\right) &= \Psi^s\left(p_{i,c}\right) \\ \Psi_h\left(p_{i,c}, p_{i,h}\right) &= \Psi^d\left(p_{i,c}, p_{i,h}\right) \ \forall h \neq c \end{aligned} \tag{2.17}$$

Assumption 3 further reduces the correction functions to only two. In particular, the first correction function applies to the case in which household i considers investing in a stock that is headquartered in the same city in which it resides, while the second correction function applies to the case in which the stock's headquarters are located in a different city.[18] The resulting two correction functions, $\Psi^s(\cdot)$ and $\Psi^d(\cdot)$, can be flexibly estimated through a polynomial series expansion.

[18] In short, as in Dahl (2002), Assumptions 1-3 can be thought of as exclusions restrictions on the conditional joint distribution of household i's idiosyncratic investment error, $\epsilon_{i,c,h,j}$, and its maximum order statistic, $v_{i,c}$, given its observed location utilities $\{V_{i,\ell}\}_{\ell=1}^{C}$ (or equivalently, by the monotonicity, the location probabilities $\{p_{i,\ell}\}_{\ell=1}^{C}$), so that:

$$f\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| \{V_{i,\ell}\}_{\ell=1}^{C}\right) = f\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| \{p_{i,\ell}\}_{\ell=1}^{C}\right) = \left\{ \begin{array}{c} f^s\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}\right) \text{ if } h = c \\ f^d\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}, p_{ih}\right) \text{ if } h \neq c \end{array} \right\}$$

Then, combining the above equation with Equation (12) yields that:

$$\mathbb{E}\left(\epsilon_{i,c,h,j} \middle| v_{i,c} < 0, \left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right) = \left\{ \begin{array}{c} \dfrac{\int_{-\infty}^{+\infty}\int_{-\infty}^{0} \epsilon_{i,c,h,j} f^s\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}\right) d\epsilon_{i,c,h,j} dv_{i,c}}{p_{i,c}} \equiv \Psi^s\left(p_{i,c}\right) \text{ if } h = c \\ \dfrac{\int_{-\infty}^{+\infty}\int_{-\infty}^{0} \epsilon_{i,c,h,j} f^d\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}, p_{ih}\right) d\epsilon_{i,c,h,j} dv_{i,c}}{p_{i,c}} \equiv \Psi^d\left(p_{i,c}, p_{i,h}\right) \text{ if } h \neq c \end{array} \right\}$$

2.5. Estimation

2.5.1. Location Choice Results

Table 2.3 presents the maximum likelihood estimation results of our location choice model. This table presents two conditional logit models for household location choices. Column 1 presents the full model of location choice derived above. The choice set of households consists of the 57 MSAs. The dependent variable is an indicator variable that equals one if a household resides in a specific MSA. We use 10,712 unique households to estimate our models, which gives us 610,584 observations.

The main explanatory variables are MSA characteristics that include the log of the population number in an MSA (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), and the log of the housing price index (LogHPI). We also include all pair-wise interaction terms between the MSA characteristics variables and the household demographics, e.g. LogIncome, LogAge, Managerial, SalesSvc, WhiteCollar, BlueCollar, Male, Married, Kids (shown in Panel B of Table 2.1) in our main specification in Column 1.

The coefficient estimates in Column 1 have the same predicted sign as in other location choice models (e.g. Bishop (2007)). The t-statistics indicate that there is substantial observed heterogeneity in location preferences, whereby households with certain demographics (i.e. LogIncome, LogAge, etc.) are more likely to locate in MSAs with certain characteristics (i.e. LogPop, Unemp, Incomepc, LogHPI, etc.). In other words, the pairwise interaction terms involving household characteristics and MSA demographics are highly significant.

Take first LogPop, a measure of population density or urban versus rural areas. The coefficient on LogPop×LogIncome is 0.05 and is marginally statistically significant with a t-statistic of 1.7. This means that higher income households are more likely to locate (i.e. prefer to live) in urban centers. Highly populated cities are also

more likely to attract old households, households whose job industry is managerial instead of professional (which is the base group in our regressions), and white collar households.

Next consider the MSA unemployment rate. The coefficient on Unemp×BlueCollar is 0.12 with a t-statistic of 3. So blue collar workers are more likely to be matched with MSAs whose unemployment rate is higher than average. The coefficient on Unemp×LogAge is also statistically significant probably because older households are about to or have already exited the labor force.

Our remaining two MSA characteristics are Incompc and LogHPI. Both proxy for costs of living. The coefficient on Incomepc×LogIncome is 0.05 with a t-statistic of 7. Similarly, the coefficient on LogHPI×LogIncome is 0.94 with a t-statistic of 4. The sensitivity with respect to these costs is higher for households which have low income. The same also applies for households which are older, have a family (i.e. are married or have kids) and for households which perform manual labor.

For comparison reasons, we also estimate the location choice model without the interaction between the household demographics and the MSA characteristics. The results are shown in Column 2 of Table 2.3. The fit of the model is reduced as indicated by the decreased values of both the log likelihood and the pseudo R^2. In addition, the coefficients of income per capita and house prices turn positive with a high statistical significance, loosing their interpretation as costs of living. Population also turns positive and becomes highly statistically significant. In short, it is important, as has been recognized in the location choice literature, that these pairwise interaction terms play a critical role in explaining location preferences.

2.5.2. Portfolio Choice Results

Uncorrected Results

We next estimate a portfolio choice model but without accounting for location choice. Table 2.4 presents the maximum likelihood estimation results of the portfolio choice model without correcting for location selection. The portfolio regression is run separately for every month in our sample, thus flexibly allowing for time variation in the coefficients. Standard errors are clustered at the household and MSA level, taking into account part of the unobserved heterogeneity of the decision makers and their residence cities. For compactness, in the table, we only present the monthly averages of all the coefficients along with their monthly averaged t-statistics.

Whether it stands alone in the portfolio regression or is together with household demographics, financial characteristics or demographics of the MSAs, the distance coefficient is always negative and highly statistically significant (with a t-statistic value ranging from -5 to -6). The average estimated coefficient equals -0.013 and becomes -0.014 when MSA demographics are included.

In Column 2, even after adding household demographics, the coefficient is virtually unchanged. Besides the expected negative effect of distance, the coefficients of household demographics also have the predicted sign in their estimates. In particular, investment in a stock is more likely and is increased in magnitude if a household has a high income. The coefficient on LogIncome is 0.03 with a t-statistic of 2.9. The same also occurs with age. Moreover, households with kids under the age of eighteen are less likely to invest relative to professionals (i.e. the base group).

As for the estimated coefficients of the financial characteristics, price enters negatively and size positively, with a high statistical significance, as anticipated. Households are not momentum traders and this is indicated by the sign of the respective coefficient. The estimated coefficient of volatility is not statistically significant. The

book-to-market coefficient is estimated to be positive, showing preferences for value stocks. A high turnover ratio of a stock also makes it attractive to households.

In Column 4, we additionally control for possible, relevant MSA demographics. In particular, we add the demographics of the MSA in which household resides (which we denote with subscript c) and the demographics of the MSA of the firm headquarters (which we denote with the subscript h). The estimation results show that households are less likely to own stocks located in urban areas. Importantly, we find that adding MSA demographics increases the negative value of the distance coefficient to -0.014. Most of the coefficients on the previous household demographics and stock financial characteristics also do not change much.

Based on our estimates in every month, we compute the average marginal effect of our main explanatory variable, *Distance*, on the portfolio weight. Since the Tobit model is non-linear, we first calculate the marginal effect by anchoring all of the covariates at their mean values in the sample, for each month. Then, we take the monthly average of all these marginal effects.

For our model in Column 1 of Table 2.4 (i.e. without any controls besides Distance), the economic effect of Distance on the portfolio weight, defined as the marginal effect of Distance on the portfolio weight times a 1 standard deviation increase in Distance, is -3.53 basis points or -31.5% of the average household portfolio weight on a Russell 1000 stock (which is 11.2 bps when we take into account a household's many zero stock positions). When we add in all other control variables as in Column 4 of Table 2.4, the economic effect of Distance becomes -1.85 bps or -16.5% of the average. Therefore, in line with the Local Bias summary statistics in Table 2.2, Distance is estimated to have a quite sizable economic effect on the household portfolio choice.

The complete time series evolution of the distance coefficient estimates is depicted in Figure 2.2. As years went by, the absolute value of the distance coefficient is

decreased. Our sample is short, so it is difficult to make anything out of this trend.[19]

Corrected Results

In Tables 2.5 and 2.6, we present the maximum likelihood estimation results after correcting for location selection. For exposition purposes, we first start with the simplest specification according to which portfolio weights are solely (nonlinearly) regressed on distance and the constant, α. Such a specification allows us to clearly highlight the decreases in the absolute value of the distance coefficient as one attempts to increase the order of the polynomials approximating the correction functions. Indeed, Table 2.5 shows that, even with two linear correction functions in Column 1, we achieve a 15% reduction. Recall that the distance coefficient was originally -0.013 with a t-statistic of -5.39. The coefficient is now -0.011 with a t-statistic of -4.51.[20] Notice that this reduction corresponds to the fact that the imputed location probabilities as captured by the correction function both predict a positive portfolio weight. In other words, the higher the probability of a household locating to a given MSA, the higher are the chances of that household owning stocks which are headquartered in that MSA.

This reduction is furthered increased in Column 2, once we increase the order of the polynomial approximation. The coefficient becomes -0.009 with a t-statistic of -4.56. As Column 3 of the table shows, with a 4th order polynomial the reduction is almost 50% (45% to be exact). The coefficient is now -0.006 with a t-statistic of

[19] It might possibly be driven by the technological improvements in the transmission of information across cities through the Internet (e.g. Barber and Odean (2002)), which naturally mitigates the cognitive and information costs that distance captures. Alternatively, it might be temporarily reflecting the stock-market mania of the mid-nineties. In particular, the rise of Internet stocks during this period may also be affecting the time series variation of this coefficient.

[20] The standard errors of the portfolio choice parameters might not be exact due to the imputation of the location probabilities from the first-stage estimation of the location choice model. This is true in any two-step estimation procedure. Of course, the estimated probabilities are consistent and the sample size in the conditional logit is quite large (10,712 households $\times$ 57 MSAs). Moreover, we run the investment model separately for every month, having different households in every period. The spirit of this exercise resembles the one of a bootstrap. In any case, bootstrapped standard errors for specific cross-sections are available upon request.

−5.32.

In Table 2.6, the same percentage decreases in the distance coefficient also apply, when we control for household demographics, financial characteristics and characteristics of the MSAs. Take the first three columns. The baseline portfolio weight regression now includes, in addition to distance, the household demographics. As we move from Column 1 to Column 3, we increase the order of the polynomial function as we did in the previous table. We see that the distance coefficient in Column 3 with the 4th order polynomial is −0.006, similar to before.

In Columns 4 to 6, we include both household demographics and stock characteristics in the portfolio weight regression. Again, in Column 6 with the 4th order polynomial, the coefficient is −0.007, which is similar to the previous −0.006 coefficient. In Columns 7-9, we additionally include the MSA characteristics in the portfolio weights regression. In Column 9 with the 4th order polynomial, the coefficient is −0.007. Thus, in all cases, with a 4th order polynomial approximation of the correction function, we estimate a roughly 50% reduction on the distance coefficient.

The changes in the economic effects (corresponding a 4th order polynomial approximation of the correction functions) are depicted in Table 2.7. In the absences of controls, the economic effect of distance drops to −1.66 bps or −14.8% of the mean. This decrease is about 50% from the uncorrected case (where the economic effect was -14.8% of the average). With full controls and correction for location selection, the economic effect of distance becomes -0.94 bps or −8.4% of the average. That is again, the reduction is roughly 50%.

2.5.3. Robustness

In the Appendix of Chapter 2, we conduct a number of robustness checks and extensions. First, we consider alternative distance measures. Second, we consider altenrative location choice models. Third, rather than use actual portfolio weights as the

dependent variable of interest which requires a Tobit model to capturing the censuring at zero, we consider value-weighted portfolio deviations from a market benchmark. We can use OLS regressions instead. Fourth, we extend our sample to Russell 3000 stocks, expand our sample of MSAs to over 80, and introduce additional household demographics with more detail on occupations.

2.6. A Decomposition of Local Bias

Our reduced form analysis suggests a natural model of local bias where household portfolio choice depends on expectations of stock-payoffs based on priors, which are correlated with subjective latent expectations in the locational choices, and i.i.d. private signals about where the precision of these signals is increasing with proximity to firm headquarter in line with the familiarity heuristic. We can then estimate this model using the residuals from the location choice model as proxies for these priors.

2.6.1. Model

Specifically, we assume that retail investors are *risk neutral* and have *subjective* beliefs about the one-period ahead returns of stocks. We also assume the existence of a risk-free asset. Investor i has a normal prior for stock j's excess return, $\tilde{f}_j \equiv \tilde{R}_j - R_f$, based on stock j's financial characteristics, $\boldsymbol{X}_j$, and his expectation about stock j's headquarters' city, $L_{i,j}$:

$$\tilde{f}_j \,|\boldsymbol{X}_j \sim \mathcal{N}\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j}, \frac{1}{\tau^0}\right) \tag{2.18}$$

Investor i receives a private signal for stock j's return:

$$S_{i,j} = \tilde{f}_j + \eta_{i,j} \tag{2.19}$$

where $\eta_{i,j} | \boldsymbol{X}_j, L_{i,j}, Dist_{i,j} \sim \mathcal{N}(0, \tau^{\eta}(Dist_{i,j}))$. That is, conditional on stock j's risk factors, $\boldsymbol{X}_j$, investor i's expectation about stock j's headquarters' city, $L_{i,j}$ and the distance between investor i's residence and stock j's headquarters, $Dist_{i,j}$, the signal's noise, $\eta_{i,j}$, is assumed to be normally distributed, with zero mean and precision $\tau^{\eta}(Dist_{i,j})$. The precision is expected to *decrease* with the distance (i.e., $\tau^{\eta}(Dist_{i,j})' < 0$). Consequently, by the projection theorem, investor i's updated subjective expectation for stock j's excess return is:

$$\mathbb{E}\left(\tilde{f}_j | \boldsymbol{X}_j, L_{i,j}, D_{i,j}, S_{i,j}\right) = \alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j} \tag{2.20}$$

where, since the private signal, $S_{i,j}$, is unobservable to the econometrician, we define investor i's *latent demand* for stock j as:

$$\xi_{i,j} \equiv \frac{\tau^{\eta}(Dist_{i,j})}{\tau^0 + \tau^{\eta}(Dist_{i,j})} \left[S_{i,j} - (a + \boldsymbol{b}\boldsymbol{X}_j + \nu L_{i,j})\right] \tag{2.21}$$

The latent demand, $\xi_{i,j}$, reflects investor i's private information about stock j as well as his optimism or pessimism about the stock's prospects. Equations (2.18) and (2.19) imply that:

$$\xi_{i,j} | \boldsymbol{X}_j, L_{i,j}, Dist_{i,j} \sim \mathcal{N}\left(0, \underbrace{\frac{\tau^{\eta}(Dist_{i,j})}{\tau^0\left[\tau^0 + \tau^{\eta}(Dist_{i,j})\right]}}_{\sigma^2_{i,j}}\right) \tag{2.22}$$

so that the variance of investor i's latent demand for stock j (namely, $\sigma^2_{i,j}$) is expected to decrease with the distance between investor i's residence and stock j's headquarters (given that the precision of the signal is conjectured to decrease as well).[21]

By the risk neutrality, Equation (2.20) is the only relevant moment of stock j's return for investor i's objective, which is to maximize the expected excess return of

[21] That is, $\frac{d\sigma^2_{i,j}(Dist_{i,j})}{dDist_{i,j}} = \frac{\tau^{\eta}(Dist_{i,j})'}{[\tau^0 + \tau^{\eta}(Dist_{i,j})]^2} < 0$, if $\tau^{\eta}(Dist_{i,j})' < 0$.

his wealth under the presence of short-selling constraints and trading costs. Indeed, households do not short, so that investor i's portfolio weight on a stock cannot be negative, i.e. $\{w_{i,j} \geq 0\}_{j \in J}$. Moreover, we assume that investors face quadratic costs from transactions. We specify the quadradicity of transaction costs in terms of acquired market value, so that when investor i acquires $n_{i,j}$ shares of stock j, he pays:

$$TC_{i,j} = \frac{1}{2}\Lambda_i \left(P_j n_{i,j}\right)^2 \tag{2.23}$$

where Λ_i scales the level of transaction cost that investor i faces for every unit of money spent on stock j. Conveniently, we let that scaling factor be:

$$\Lambda_i = \frac{c}{W_i^2} \tag{2.24}$$

The more wealth an investor has, the less he has to worry about transaction costs. If the dependence on investor i's wealth, W_i, is quadratic, transaction costs matter much less for richer households and essentially refer to the level of portfolio weights.

Combining Equations (2.20), (2.23) and (2.24), we end up expressing investor i's objective as follows:[22]

$$\max_{\{w_{i,j} \geq 0\}_{j \in \mathcal{J}}} \left\{ \sum_{j \in \mathcal{J}} \left[\left(\alpha + \boldsymbol{\beta} \boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j}\right) w_{i,j} - \frac{1}{2} c w_{i,j}^2 \right] \right\} \tag{2.25}$$

The KKT conditions then imply that:

$$w_{i,j} = \left(\frac{\alpha + \boldsymbol{\beta} \boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j}}{c} \right)^+ \tag{2.26}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ simultaneously captures investor i's decision of whether and how much to invest in every stock included in his consideration set. According to

[22] Investor i's budget constraint requires the sum of his stock-portfolio weights to equal one minus the weight on the risk-free asset. The latter is assumed to be perfectly adjustable to the needs of the portfolio optimization problem.

Equation (2.26), investor i invests in stock j provided that the expected excess return of the stock is positive in his view. The transaction cost to which he is subject determines the actual level of the portfolio weight, scaling it up or down.

Since from Equation (2.22) the latent demand, $\xi_{i,j}$, is normally distributed, Equation (2.26) constitutes a heteroskedastic tobit model. However, as the model stands, it is under-identified. To this end, we make the following simplifications; First, we normalize investors' precision, τ^0 and transaction cost parameter, c, to be equal to 1. And second, we parameterize the precision of the signal's noise, $\eta_{i,j}$, as follows:

$$\tau^{\eta}(Dist_{i,j}) = \frac{\exp[2\left(\gamma + \delta Dist_{i,j}\right)]}{1 - \exp[2\left(\gamma + \delta Dist_{i,j}\right)]} \tag{2.27}$$

Equation (2.22) together with (2.26) entails that:

$$w_{i,j} = \left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \exp\left(\gamma + \delta Dist_{i,j}\right)\zeta_{i,j}\right)^{+} \tag{2.28}$$

where $\zeta_{i,j}\,|\boldsymbol{X}_j, L_{i,j}, Dist_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Equation (2.28) is the standard textbook expression for a tobit model with heteroskedasticity (e.g. Wooldridge (2010)). Here, the heteroskedasticity is specified in terms of the distance of investor i's residence from stock j's headquarters' city.

As a proxy for investor i's priors about stock j's headquarters' city, we use the Pearson residuals from the conditional logit model in Section 2. That is, we define:

$$L_{i,j} \equiv \frac{r_{i,j} - p_{i,j}}{\sqrt{p_{i,j}}} \tag{2.29}$$

where $r_{i,j}$ equals 1 if investor i resides in stock j's headquarters' city and 0 otherwise, while $p_{i,j}$ is the investor i's predicted location probability in that city.

2.6.2. Estimation

We estimate the model for every month in our sample separately. For comparison, we also estimate two restricted versions that the model nests (henceforth referred to as Full Model). In the first version, the coefficient of investor i's priors about stock j's headquarters' city, $L_{i,j}$, equals zero (henceforth referred to as No Residual). In the second version, the coefficient of the distance between investor i's residence and stock j's headquarters, $Dist_{i,j}$, equals zero (henceforth referred to as No Distance). We present the time series averages of the coefficients in Panel A of Table 2.8 along with the average t-statistics [in brackets], based on standard errors calculated using the estimated asymptotic variance-covariance matrix of the maximum-likelihood estimator (i.e. the inverse of the information matrix evaluated at the maximum-likelihood estimates).

As reported in Column 1 of Panel A, the estimated coefficient of Residual is on average 0.118, with an average t-statistic of 43.8. On the other hand, the coefficient of Distance is estimated on average to equal -0.002, with an average t-statistic of approximately -13. In Column 2, where the coefficient of Residual is restricted to be zero, the average estimated coefficient of Distance is -0.004 (i.e., two times larger in magnitude than in the Full Model), with an average t-statistic of -30.7. In Column 3, where the coefficient of Distance is constrained at zero, Residual has an average estimated coefficient equal to 0.13 (i.e., about 10% higher than in the Full Model), with an average t-statistic of 50.6.

Turning to the coefficients of the stock characteristics, their estimates are very similar across the three different models. Regarding the estimates in the Full Model, which are depicted in Column 1, the coefficient of LogSize equals on average 0.32, with an average t-statistic of 63.1. The average estimated coefficient of BTM is 0.08, with an average t-statistic of 6.7. Turnover has an average coefficient of 0.361, with an average t-statistic of 27.98. Momentum has an average coefficient of -0.148 with

an average t-statistic of -12.0, while Volatility has an average coefficient of 0.344 with an average t-statistic of 9.8.

Panel B of Table 2.8 depicts the economic effects of Residual and Distance. In Column 1 (which corresponds to the estimates in the Full Model), the estimated economic effect of Residual is on average 1.61 bps or 14.4% of the mean portfolio weight on a stock in Russell 1000. In the same column, the economic effect of Distance is on average -1.03 bps or -9.2% of the mean. Hence, the economic effect of Residual (which expresses the latent expectations from the locational decisions)is bigger than the economic effect of Distance. Household priors account for 60% of local bias and familiarity or distance 40%.

In Column 2 (which refers to the estimates in the No Residual Model), the economic effect of Distance equals on average -2.45 bps or -21.9% of the mean. That is, its magnitude is about 138% larger than in the Full Model. In Column 3 (which corresponds to the results in the No Distance Model), the economic effect of Residual is on average 1.81 bps or 16.1% of the mean. This figure is 12.4% higher than the economic effect of Residual in the Full Model.

2.6.3. Model Fit

According to the estimation results of the Full Model, the coefficients of Residual and Distance are statistically significant. In this subsection, we also perform a nested likelihood ratio tests for hypotheses according to which only one of the two coefficients is statistically significant. The results of these tests are presented in Table 2.9.

In the first row of the table, we test the null hypothesis $H_0 : \nu = 0$ (No Residual Model) against the alternative hypothesis $H_1 : \nu \neq 0$ (Full Model). Under the null, investor i's priors about stock j's headquarters' city, $L_{i,j}$, does not affect his portfolio weight on the stock. The test is performed by calculating the test statistic $D = 2[\log(L_{H_1}) - \log(L_{H_0})]$ (i.e., twice the difference between the log-likelihood in the

Full Model and the log-likelihood in the No Residual Model) in every month. We then average over time the monthly test statistics to obtain the value of the D-statistic in respective sample period. The calculated value is remarkably high (i.e., 1913.6). To test H_0, we compare this figure to critical values of the χ^2 distribution with 1 degree of freedom. We reject the null hypothesis at any reasonable value of statistical significance, confirming that investors' latent expectations about the cities in which they can locate affect their subsequent investment decisions.

In the same spirit, we test the null hypothesis $H_0 : \delta = 0$ (No Distance Model) against the alternative $H_1 : \delta \neq 0$ (Full Model) in the second row of the table. According to this null, the distance investor i's residence and stock j's headquarters' city, $Distance$, does not affect investor i's precision regarding stock j's signal. The value of the D-statistic is now approximately 174.5. This figure is again much higher than critical values of the χ^2 distribution with 1 degree of freedom, at any reasonable significance level. Hence, H_0 is rejected, showing that geographical proximity has a statistically significant effect on the investor's portfolio choices.

2.7. Conclusion

This local bias puzzle is generally explained by theories that assign a causal role to proximity. The empirical analyses typically assume that households locate randomly. But a household in practice optimally locates in a city depending latent subjective expectations about the economic prospects of a city, which correlated with demand for local stocks. We propose a correction for this selection bias in reduced-form portfolio regression using location choice models. We then propose and estimate a structural model of local bias driven by household priors about stocks in a city, using location-choice model residuals, and familiarity, whereby confidence about stock-payoff signals rise with proximity.

Table 2.1: Summary Statistics

This table reports the summary statistics of all the variables in our sample. Panel A shows the characteristics of the 57 Metropolitan Statistical Areas (MSAs). Pop is the population number, LogPop is the log of population, Unemp is the unemployment rate, Incomepc is the income per capita, HPI is the housing price index, and LogHPI is the log of HPI. Panel B shows the demographics of the households in our sample. Income is the income level. LogIncome is the log of Income. Age is the age of the household head. LogAge is the log of Age. Professional, Managerial, SalesSvc, WhiteCollar, BlueCollar, Male, Married are dummy variables that equal one if the household head has a professional-type job, managerial-type job, sales-services-type job, white collar-type job, blue collar-type job, is a male, and is married respectively. Kids is the number of kids under the age of 18 in a household. Panel C refers to the stock holdings of the households and their distance. Portwt is the portfolio of a household on a stock. Numst is the number of stocks that a household holds. Distance is the distance between a household's residential zip code area and the zip code area of a stock's headquarters. Away250m (Away100m) is a dummy indicating whether a stock is headquartered more than 250 miles (100 miles) away from a household's residence. Panel D shows the characteristics of the Russell 1000 stocks. Price is the price of a stock. LogPrice is the log of Price. Size is the market capitalization. LogSize is the log of Size. BTM is the book-to-market ratio. Turnover is share turnover. Momentum is the past 12-month return. Volatility is the volatility of the monthly returns in the past 12 months. The last 17 variables of Food to Other are industry dummy variables that equal to one if a stock belongs to that particular industry. The 17 industries are defined based on the categorization of the Fama-French 17-industry portfolios. The sample is from January 1991 to November 1996.

Panel A: Characteristics of 58 Metropolitan Statistical Areas (MSAs)

	Mean	S.D.	Median	Min	Max
Pop (million)	2.50	2.80	1.50	0.68	18.00
LogPop	14.43	0.71	14.25	13.43	16.69
Unemp (%)	5.80	1.92	5.49	2.92	15.63
Incomepc ($ thousand)	23.62	3.91	23.06	16.96	44.88
HPI (=100 in 1995Q1)	99.19	7.85	100.37	64.57	123.15
LogHPI	4.59	0.08	4.61	4.17	4.81

Panel B: Demographics of Households

	Mean	S.D.	Median	Min	Max
Income ($ thousand)	82.40	31.69	87.50	10.00	125.00
LogIncome	11.22	0.50	11.38	9.21	11.74
Age (years)	47.73	10.59	46.00	24.00	94.00
LogAge	3.84	0.22	3.83	3.18	4.54
Professional (%)	55.1				
Managerial (%)	27.7				
SalesSvc (%)	8.4				
WhiteCollar (%)	4.9				
BlueCollar (%)	3.9				
Male (%)	92.6				
Married (%)	79.2				
Kids	0.67	0.94	0	0	5

Panel C: Household Stock Holdings

	Mean	S.D.	Median	Min	Max
Portval ($ thousand)	39.84	152.10	14.72	0.00	5,108.17
Portwt	11.2bps	0.03	0	0	1
Numst	1.85	1.55	1	1	24
Distance (degrees)	17.60	11.45	15.66	0.00	39.72
Away250m (%)	88.2				
Away100m (%)	93.6				

Table 2.1 Cont'd: Summary Statistics

Panel D: Characteristics of Russell 1000 Stocks					
	Mean	S.D.	Median	Min	Max
Price ($)	33.94	25.06	29.62	0.03	684.00
LogPrice ($)	3.32	0.69	3.39	-3.51	6.53
Size ($ million)	3455.53	8106.10	1229.75	0.56	130000.00
LogSize ($ million)	7.23	1.25	7.11	-0.57	11.78
BTM	0.43	4.09	0.43	-271.51	320.13
Turnover	0.30	0.93	0.17	0.00	113.05
Momentum	0.17	2.15	0.07	-0.99	217.09
Volatility	0.18	1.01	0.12	0.00	67.58
Food (%)	4.0				
Mines (%)	2.0				
Oil (%)	6.0				
Clths (%)	1.0				
Durbl (%)	1.0				
Chems (%)	3.0				
Cnsum (%)	5.0				
Cnstr (%)	2.0				
Steel (%)	2.0				
FabPr (%)	1.0				
Machn (%)	12.0				
Cars (%)	2.0				
Trans (%)	3.0				
Utils (%)	6.0				
Rtail (%)	8.0				
Finan (%)	16.0				
Other (%)	26.0				

Table 2.2: Local Bias among Households

This table provides summary statistics for the local bias of household portfolio holdings in our data. Column 2 (Avg. Distance from Holdings) reports the average distance of households from the stocks they hold in their portfolios. The average distance in Column 2 is computed as $\frac{1}{I}\sum_i \sum_j w_j^i d_j^i$, where d_j^i is the distance between household i's residential area and the headquarters area of stock j, w_j^i is household i's portfolio weight on stock j, and I is the total number of households. Column 3 (Avg. Distance from Benchmark) is computed as $\frac{1}{I}\sum_i \sum_j \bar{w}_j d_j^i$, where $\bar{w}_j$ is a benchmark Russell 1000 portfolio weight on stock j, and d_j^i and I are the same as in Column 2. In Row 1 the benchmark is the equally weighted portfolio. In Row 2 the benchmark is the value-weighted portfolio. Column 4 (Difference) reports the difference between Column 2 and Column 3, which is the local bias in distance units. Column 5 (% Bias (LB)) reports the local bias (LB) measure as a percentage. Column 6 (t-stat) reports the t-statistics for the LB measure. The sample period is from January 1991 to November 1996.

	Avg. Distance from				
Weights	Holdings	Benchmark	Difference	% Bias (LB)	t-stat
Equal	9.38	17.67	8.29	45.45	52.74
Value	9.38	17.65	8.26	43.72	47.65

Table 2.3: Conditional Logit Estimation of Household Location Choice

This table presents the results from the maximum likelihood estimation of the two conditional logit models for household location choices. t-statistics [shown in brackets] are based on standard errors clustered at the household level. The choice set of households consists of 57 MSAs. The dependent variable is an indicator variable that equals one if a household resides in the specific MSA. The main explanatory variables are MSA characteristics that include the log of the population number in an MSA (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), and the log of the housing price index (LogHPI). We also include all pair-wise interaction terms between these MSA characteristics variables and the household demographics variables LogIncome, LogAge, Managerial, SalesSvc, WhiteCollar, BlueCollar, Male, Married, Kids (shown in Panel B of Table 2.1) in our main location choice model in Column 1. For comparison, we show the estimation results from the model where the interaction terms are not included in Column 2. Log-likelihood is the estimated log-likelihood value. Pseudo R^2 is the McFadden's pseudo R^2 measure based on the estimated log-likelihood value. I is the number of households.

	(1)	(2)
LogPop	-0.265	0.938
	[-0.64]	[71.12]
Unemp	-0.418	-0.124
	[-1.51]	[-14.17]
Incomepc	-0.240	0.075
	[-2.42]	[24.00]
LogHPI	-3.712	0.507
	[-1.12]	[4.73]
LogPop×LogIncome	0.051	
	[1.70]	
LogPop×LogAge	0.156	
	[2.48]	
LogPop×Managerial	0.096	
	[3.07]	
LogPop×SalesSvc	0.039	
	[0.76]	
LogPop×WhiteCollar	0.210	
	[3.19]	
LogPop×BlueCollar	0.093	
	[1.28]	
LogPop×Male	0.056	
	[1.15]	
LogPop×Married	-0.064	
	[-1.83]	
LogPop×Kids	-0.012	
	[-0.79]	
Unemp×LogIncome	-0.008	
	[-0.44]	
Unemp×LogAge	0.094	
	[2.18]	
Unemp×Managerial	0.003	
	[0.15]	
Unemp×SalesSvc	-0.006	
	[-0.17]	
Unemp×WhiteCollar	0.058	
	[1.41]	
Unemp×BlueCollar	0.118	
	[3.04]	

Table 2.3 Cont'd: Conditional Logit Estimation of Household Location Choice

	(1)	(2)
Unemp×Male	0.013	
	[0.34]	
Unemp×Married	-0.010	
	[-0.42]	
Unemp×Kids	0.017	
	[1.72]	
Incomepc×LogIncome	0.051	
	[6.96]	
Incomepc×LogAge	-0.045	
	[-3.06]	
Incomepc×Managerial	-0.009	
	[-1.24]	
Incomepc×SalesSvc	-0.018	
	[-1.48]	
Incomepc×WhiteCollar	-0.002	
	[-0.09]	
Incomepc×BlueCollar	0.003	
	[0.16]	
Incomepc×Male	-0.051	
	[-4.89]	
Incomepc×Married	-0.038	
	[-4.64]	
Incomepc×Kids	0.000	
	[0.08]	
LogHPI×LogIncome	0.936	
	[3.96]	
LogHPI×LogAge	-1.186	
	[-2.32]	
LogHPI×Managerial	-0.269	
	[-1.09]	
LogHPI×SalesSvc	-0.333	
	[-0.86]	
LogHPI×WhiteCollar	0.253	
	[0.46]	
LogHPI×BlueCollar	-1.258	
	[-2.22]	
LogHPI×Male	-0.777	
	[-1.67]	
LogHPI×Married	-0.607	
	[-2.05]	
LogHPI×Kids	-0.516	
	[-4.20]	
Log-likelihood	-38,579.5	-38,747.5
Pseudo R^2	0.109	0.105
I	10,712	10,712

Table 2.4: Tobit Estimation of Household Portfolio Choice, No Correction for Location Selection

This table presents the estimation results from four Tobit models of household portfolio choice without correcting for location selection. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographics controls are: the log of a household's income (LogIncome), the log of the age of a household head (LogAge), fixed effects for the occupation of a household's head (Managerial to BlueCollar), gender (Male) and marital status (Married), and the number of kids under the age of 18 a household has (Kids). The stock characteristics controls are: the log of the price (LogPrice), the log of the market capitalization (LogSize), the book-to-market ratio (BTM), the turnover ratio (Turnover), the momentum of the past 12 months (Momentum), the volatility (Volatility) and industry fixed effects based on the 17 Fama-French industry portfolios. Controls for MSA demographics include: the log of the population number (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), the log of the housing price index (LogHPI), where the subscripts c and h denote respectively the MSAs of household i's residence and stock j's headquarters. The constant term and the standard deviation of the normal error term in the Tobit regression are not reported for brevity. Panel B presents the average marginal effects of Distance on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is January 1991 to November 1996.

	Dependent Variable: Portfolio Weight $w_{i,j}$			
	(1) No Controls	(2) + HH Demo's	(3) + Stock Char's	(4) +MSA Demo's
Distance	-0.013	-0.013	-0.013	-0.014
	[-5.39]	[-5.39]	[-6.12]	[-6.64]
LogIncome		0.035	0.03	0.028
		[3.27]	[2.81]	[2.65]
LogAge		0.084	0.09	0.096
		[4.37]	[4.7]	[4.62]
Male		0.021	0.025	0.028
		[1.22]	[1.59]	[1.93]
Married		-0.017	-0.016	-0.015
		[-1.58]	[-1.53]	[-1.33]
Kids		-0.014	-0.014	-0.013
		[-2.16]	[-2.35]	[-2.23]
LogPrice			-0.376	-0.381
			[-19.98]	[-19.97]
LogSize			0.429	0.429
			[39.57]	[37.71]
BTM			0.099	0.1
			[4.62]	[4.7]
Turnover			0.396	0.377
			[18.49]	[20]
Momentum			-0.035	-0.036
			[-1.39]	[-1.29]
Volatility			0.02	0.004
			[0.59]	[0.22]
LogPop_c				-0.02
				[-1.61]
Unemp_c				0.001
				[0.12]
Incomepc_c				0.002
				[0.58]
LogHPI_c				-0.356
				[0.35]
LogPop_h				-0.029
				[-1.66]
Unemp_h				0.01
				[0.75]
IncPerCap_h				0.004
				[1.21]
LogHPI_h				-0.717
				[-0.1]
HH Occupation F.E.	NO	YES	YES	YES
Stock Industry F.E.	NO	NO	YES	YES

Table 2.5: Tobit Estimation of Household Portfolio Choice with Correction for Location Selection, No Controls

This table presents the estimation results from three Tobit models of household portfolio choice in the presence of local bias, without controls but correcting for location selection. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and city level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. The correction functions $\Psi^s(p_{i,c})$ and $\Psi^d(p_{i,h}, p_{i,c})$ for location selections are approximated by polynomials. The approximation is linear (in Column 1), quadratic (in Column 2) and quartic (in Column 3). The constant term and the standard deviation of the normal error term in the Tobit regression are not reported for brevity. Panel B presents the average marginal effects on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is January 1991 to November 1996.

	Dependent Variable: $w_{i,j}$		
	(1)	(2)	(3)
	Linear Ψ's	Quadratic Ψ's	Quartic Ψ's
Distance	-0.011	-0.009	-0.006
	[-4.51]	[-4.56]	[-5.32]
p_c^s	2.832	14.859	61.585
	[2.58]	[3.55]	[7.43]
p_h^d	0.895	-0.44	8.712
	[3.28]	[-0.42]	[2.99]
p_c^d	-0.89	-3.49	-18.759
	[-2.84]	[-2.84]	[-55.24]
$(p_c^s)^2$		-0.922	-14.198
		[-2.82]	[-468.16]
$(p_h^d)^2$		0.045	-2.686
		[0.78]	[-4.73]
$(p_c^d)^2$		0.133	2.951
		[2.11]	[37.59]
$p_h^d p_c^d$		0.299	3.874
		[5.27]	[7.38]
$(p_c^s)^3$			103.538
			[6.12]
$(p_h^d)^3$			23.603
			[5.23]
$(p_c^d)^3$			-18.474
			[-36.83]
$(p_h^d)^2 p_c^d$			-0.269
			[-6.58]
$p_h^d (p_c^d)^2$			-0.247
			[-8.29]
$(p_c^s)^4$			-236.961
			[-5.62]
$(p_h^d)^4$			-61.354
			[-5.08]
$(p_c^d)^4$			39.312
			[52.64]
$(p_h^d)^3 p_c^d$			0.637
			[4.3]
$(p_h^d)^2 (p_c^d)^2$			0.006
			[5.61]
$p_h^d (p_c^d)^3$			0.489
			[6.02]

Table 2.6: Tobit Estimation of Household Portfolio Choice with Correction for Location Selection, Full Controls

This table presents the estimation results from nine Tobit models of household portfolio choice with controls, that correct for location selection using polynomial approximations for the correction functions. In Columns 1, 4 and 7, the approximation of the correction functions is linear. In Columns 2, 6 and 8, the approximation of the correction functions is quadratic. In Columns 3, 7 and 9, the approximation of the correction functions is quartic. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. In Columns 1, 2 and 3, household demographics are included as controls. In Columns 4, 5 and 6, stock characteristics are added as controls. In Columns 7, 8 and 9, MSA demographics are added as controls. The household demographics controls are: the log of a household's income (LogIncome), the log of the age of a household head (LogAge), fixed effects for the occupation of a household's head (Managerial to BlueCollar), gender (Male) and marital status (Married), and the number of kids under the age of 18 a household has (Kids). The stock characteristics controls are: the log of the price (LogPrice), the log of the market capitalization (LogSize), the book-to-market ratio (BTM), the turnover ratio (Turnover), the momentum of the past 12 months (Momentum), the volatility (Volatility) and industry fixed effects based on the 17 Fama-French industry portfolios. Controls for MSA demographics include: the log of the population number (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), the log of the housing price index (LogHPI), where the subscripts c and h denote respectively the MSAs of household i's residence and stock j's headquarters. The constant term and the standard deviation of the normal error term in the Tobit regression are not reported for brevity. The sample period is January 1991 to November 1996.

	Dependent Variable: Portfolio Weight $w_{i,j}$								
	(1) Lin.	(2) Quad.	(3) Quart.	(4) Lin.	(5) Quad.	(6) Quart.	(7) Lin.	(8) Quad.	(9) Quart.
	+ HH Demo's			+ Stock Char's			+MSA Demo's		
Distance	-0.011	-0.009	-0.006	-0.011	-0.009	-0.007	-0.012	-0.009	-0.007
	[-4.52]	[-4.58]	[-6.13]	[-5.86]	[-6.43]	[-8.2]	[-6.32]	[-7.3]	[-10.26]
LogIncome	0.032	0.031	0.02	0.039	0.032	0.024	0.042	0.03	0.019
	[2.87]	[2.35]	[2.31]	[3.49]	[2.39]	[2.7]	[3.32]	[2.24]	[1.79]
LogAge	0.082	0.089	0.089	0.094	0.1	0.105	0.101	0.102	0.105
	[4.1]	[4.38]	[4.95]	[4.7]	[4.79]	[6.53]	[4.86]	[4.95]	[5.2]
Male	0.021	0.025	0.022	0.021	0.027	0.027	0.022	0.031	0.03
	[1.25]	[1.55]	[1.69]	[1.37]	[1.85]	[2.2]	[1.45]	[2.22]	[2.46]
Married	-0.013	-0.011	-0.001	-0.023	-0.015	-0.008	-0.027	-0.012	-0.004
	[-1.12]	[-0.89]	[-0.14]	[-1.98]	[-1.23]	[-0.83]	[-2.28]	[-1.01]	[-0.5]
Kids	-0.014	-0.013	-0.013	-0.015	-0.013	-0.013	-0.014	-0.012	-0.012
	[-2.09]	[-2.03]	[-2.43]	[-2.41]	[-2.29]	[-2.45]	[-2.39]	[-2.12]	[-2.28]
LogPrice				-0.376	-0.375	-0.37	-0.379	-0.378	-0.374
				[-20.01]	[-20.35]	[-24.76]	[-19.45]	[-20.17]	[-25.51]
LogSize				0.43	0.428	0.423	0.429	0.428	0.426
				[39.65]	[40.19]	[50.95]	[38.47]	[41.22]	[49.69]
BTM				0.101	0.101	0.102	0.101	0.101	0.102
				[4.74]	[4.75]	[5.65]	[4.81]	[4.82]	[5.56]
Turnover				0.391	0.385	0.37	0.371	0.368	0.356
				[19.26]	[22.93]	[24.25]	[20.85]	[23.01]	[24.59]
Momentum				-0.036	-0.035	-0.035	-0.037	-0.036	-0.036
				[-1.36]	[-1.35]	[-1.64]	[-1.29]	[-1.26]	[-1.58]
Volatility				0.02	0.024	0.017	0.004	0.006	0.001
				[0.54]	[0.68]	[0.56]	[0.19]	[0.23]	[0.25]
LogPop_c							0.006	-0.014	-0.04
							[0.33]	[-0.66]	[-1.99]
Unemp_c							-0.002	-0.003	-0.003
							[-0.24]	[-0.55]	[-0.46]
Incomepc_c							0.003	0.001	-0.004
							[0.84]	[0.07]	[-6.59]
LogHPI_c							-0.303	-0.199	-0.047
							[0.5]	[-0.15]	[-6.62]
LogPop_h							0.008	0.013	-0.11
							[0.09]	[-0.08]	[-5.28]
Unemp_h							0.01	0.007	0.023
							[0.82]	[1.03]	[2.93]
Incomepc_h							0.007	0.006	-0.005
							[1.92]	[1.51]	[-14.29]
LogHPI_h							-0.552	-0.441	0.016
							[0.11]	[-0.78]	[-1.22]
HH Occupation F.E.	YES	YES	YES	YES	YES	YES	YES	YES	YES
Stock Industry F.E.	NO	NO	NO	YES	YES	YES	YES	YES	YES

Table 2.7: Economic Effects of Distance on Household Portfolio Weights

This table shows the economic effect of a one standard deviation increase in Distance on the portfolio weight of a household on a stock, from different specifications of our household portfolio choice model. The economic effect is calculated based on the marginal effect of Distance on the portfolio weight from the Tobit estimation, with all other control variables anchored at their mean levels. The marginal effect is computed in each monthly estimation and then averaged across all months. "Uncorrected" vs. "Corrected" denotes whether the 4th order correction function for location selection bias is included. "No Controls" vs. "Full Controls" denotes whether the full set of controls for household demographics, stock characteristics and MSA demographics are included. The economic effect is measured in basis points and as percentage of the average household portfolio weight on any stock in the investment universe (taking into account the extensive margin). The sample period is January 1991 to November 1996.

	No Controls	Full Controls
Uncorrected	-3.53 bps -31.5%	-1.85 bps -16.5%
Corrected	-1.66 bps -14.8%	-0.94 bps -8.4%

Table 2.8: Structural Estimation of Retail Investor Portfolio Choice

This table presents the estimation results from three structural Tobit models of retail investor portfolio choice. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors calculated using the estimated asymptotic variance-covariance matrix of the maximum-likelihood estimator, i.e. the inverse of the information matrix evaluated at the maximum-likelihood estimates. (The employed standard errors equal the square root of the diagonal terms of this matrix.) The dependent variable is $w_{i,j}$, the portfolio weight of household i on stock j. The first key explanatory variable is Residual, i.e. investor i's priors about stock j's headquarters' city measured by the Pearson residuals from our multinomial logit. The second key explanatory variable is Distance, i.e. the distance between household i's residential ZIP Code and the ZIP Code of stock j's headquarters. Stock j's characteristics are also included as controls: LogSize is the log of market capitalization, BTM is book to market ratio, Turnover is the monthly turnover ratio, Momentum is the past annual return, Volatility is the standard deviation of monthly returns in the past year. The Tobit models are heteroskedastic, with the standard deviation of the error term being equal to $\gamma \exp(\delta \cdot Distance)$. Panel B presents the average marginal effects of Residual and Distance on $w_{i,j}$ in basis points and as percentage of the mean. In Column 1, both Residual and Distance enter in the structural model. In Column 2, the coefficient of Residual is constrained to equal zero. In Column 3, the coefficient of Distance is constrained to equal zero. The sample period is from January 1991 until November 1996.

Panel A: Coefficient Estimates

	(1) Full Model	(2) No Residual	(3) No Distance
Constant	-5.953	-5.988	-5.979
	[-78.97]	[-78.85]	[-78.98]
LogSize	0.32	0.321	0.32
	[63.07]	[62.92]	[63.02]
BTM	0.08	0.079	0.081
	[6.68]	[6.56]	[6.67]
Turnover	0.361	0.39	0.354
	[27.98]	[30.31]	[27.3]
Momentum	-0.148	-0.149	-0.149
	[-12.04]	[-12.11]	[-12.05]
Volatility	0.347	0.361	0.344
	[9.85]	[10.11]	[9.77]
Residual	0.118	-	0.13
	[43.76]	[-]	[50.59]
Constant S.D.	1.195	1.255	1.168
	[88]	[87.99]	[89.15]
Distance	-0.002	-0.004	-
	[-12.96]	[-30.72]	[-]

Panel B: Economic Effects in Basis Points and as Percentage of the Mean

	(1) Whole Sample	(2) No Residual	(3) No Distance
Residual	1.61	-	1.81
	14.4%	[-]	16.1%
Distance	-1.03	-2.45	-
	-9.2%	-21.9%	[-]

Table 2.9: Nested Likelihood Ratio Test for Restricted Versions of the Structural Demand Model

This table presents nested likelihood ratio tests for shutting down the impact of either the investor i's expectation about stock j's headquarters' city or the distance between investor i's residence and stock j's headquarters. The first row refers to testing the null hypothesis $H_0 : \nu = 0$ (No Residual) against the alternative hypothesis $H_1 : \nu \neq 0$ (Full Model). Under this null, investor i's priors about stock j's headquarters' city does not affect his portfolio weight on the stock. The second row refers to testing the null hypothesis $H_0 : \delta = 0$ (No Distance) against the alternative hypothesis $H_1 : \delta \neq 0$ (Full Model). Under this null, the distance between investor i's residence and stock j's headquarters does not affect investor i's precision about stock j's signal. In every month, the test statistic $D = 2[log(L_{H_1}) - log(L_{H_0})]$ is calculated (which is twice the difference between the log-likelihood in the Full Model and the log-likelihood in the Restricted Model). The D-statistics depicted below are the time-series averages. The presented average D-statistics can be compared with the critical values of the χ^2 distribution with 1 degree of freedom, e.g. 10.83 at the $\alpha = 0.1\%$ level of statistical significance. The sample period is from January 1991 until November 1996. The universe of stocks is Russell 1000. Households live in 57 MSAs.

Hypothesis Test	Average D-Statistic
No Residual vs. Full Model	1913.57
No Distance vs. Full Model	174.52

Figure 2.1: Geographical Distribution of Households in 57 MSAs and Russell 1000 Stocks

This figure depicts the ZIP Code geographical coordinates of 10,712 households residing in 57 MSAs with a population of at least 750K at the end of the year 1996 and 1,056 publicly traded firms included in the Russell 1000 index during the sample period. The address ZIP-codes of households and the stocks' headquarters are converted to geographical coordinates based on the correspondence provided by the US Census Bureau. The horizontal axis is in longitude coordinates, while the vertical axis is in latitude coordinates. The blue circles indicate households, while the red squares indicate stocks. The sample period is from January 1991 to November 1996.

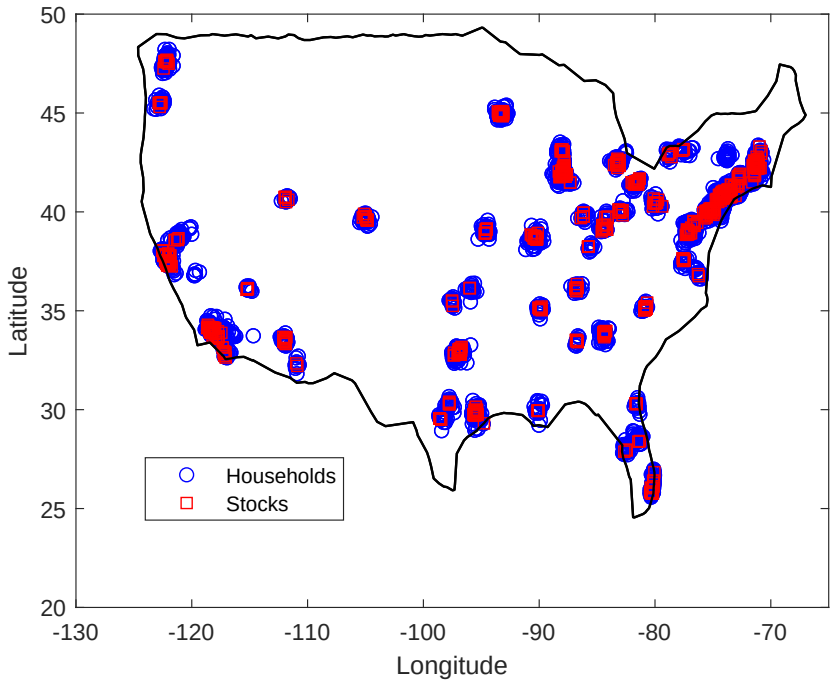
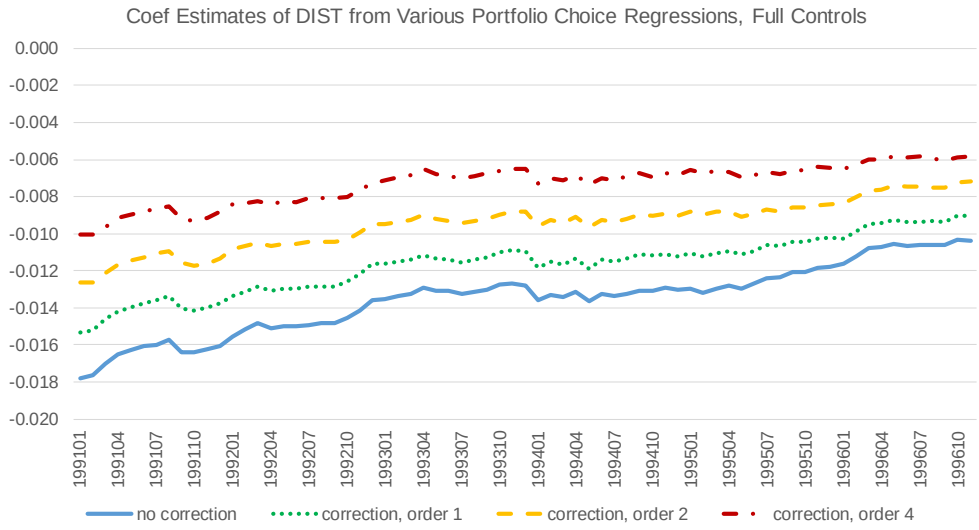


Figure 2.2: Coefficient Estimates of Distance across Time from Various Portfolio Choice Regressions with Full Controls

This figure depicts the coefficient estimates of our key explanatory variable Distance across the time periods from various portfolio choice regressions with full household demographics, stock characteristics, industry and MSA characteristics controls. The vertical axis shows the values of the estimates. The horizontal axis shows the time periods (year-month). The blue line represents the estimates from the regression without any correction for location selection biases, while the green dotted line, the orange dashed line and the red dash-dotted line represent regressions with correction functions for location selections approximated by polynomials of order 1, 2 and 4 respectively. The sample period is from January 1991 (199101) to November 1996 (199611).



Appendix of Chapter 2

In Appendix Table 2.1, we repeat the estimation of the portfolio choice model, replacing the continuous distance variable measured in degrees with dummies indicating whether the headquarters of a stock are more than a certain threshold of miles away from a household's residence. In particular, using the conservative threshold of 250 miles, the reduction in the Away coefficient is about 44%, which is similar to the one in our baseline specification (50%). Naturally, most firms have their headquarters away from a household's residence, so that the coefficient of Away is strengthened once the threshold decreases to 100 miles. Indeed, according to Panel C of Table 2.1, dropping the threshold from 250 miles to 100 miles results in a 5% increase (decrease) of distant (local) stocks. Nevertheless, even in the case of 100 miles, correcting for selection still drops the coefficient of Away by 32%.

According to our location choice model, when households locate, they match to the characteristics of a city taking into account their income and job code. That is, we allow both income and job codes to contribute to households' observed heterogeneity in locational preferences. This assumption precludes these variables from being an outcome of the location decision; we assume *perfect foresight* regarding a household's ability to earn money in any city it decides to locate and treat these variables as proxies for that ability. Yet, to alleviate any concerns regarding the plausible endogeneity of income or job codes, we repeat the whole estimation exercise excluding either income or both income and job codes from the location model. The estimation results of these two alternative location models are shown in Column 1 and 2 of Appendix Table 2.2. The estimation results of the corresponding portfolio choice models, which are presented in Appendix Table 2.3, in terms of the distance coefficient estimate and the reduction due to selection correction are virtually the same.[23]

[23]In more detail, with a fourth order polynomial approximation of the correction functions, the distance coefficient when the location choice model does not include income is -0.00714, while it

Motivated by Brandt, Santa-Clara, and Valkanov (2009), we also explore the implications of selection correction in an alternative investment model in which households deviate from value-weighted portfolios according to a linear factor model with the same variables as our baseline Tobit specification. We denote the parameters of this new model $\boldsymbol{\theta}^{dev} \equiv \left(\alpha^{dev}, \boldsymbol{\beta}^{dev}, \boldsymbol{\gamma}^{dev}, \delta^{dev}\right)$ and estimate them by running the following regression for every month in the sample period:

$$w_{i,c,h,j} - w_j^{VW} = \alpha^{dev} + \boldsymbol{\beta}^{dev}\boldsymbol{x}_j + \boldsymbol{\gamma}^{dev}\boldsymbol{D}_i + \delta^{dev} dist_{i,c,h,j} + \Psi_{c,h} + \epsilon_{i,c,h,j} \tag{2.30}$$

where the dependent variable in the LHS of the equation is the deviation of household i's portfolio weight on stock j from the value-weighted portfolio on that stock, while $\Psi_{c,h}$ denotes the corresponding correction function. The advantage of this specification over the Tobit model is that the estimate of the distance coefficient can be directly interpreted as the estimated local bias under the same null hypothesis. On the other hand, the caveat is that the many zero portfolio weights on stocks, that motivated the use of our Tobit model in the first place, are translated to many negative deviations from the market.

The estimation results of this portfolio choice model, before and after the selection adjustment, are presented in Columns 5-8 of Table 2.4. Specifically, in Column 5, without correcting for selection, the distance coefficient is about -0.43 bps. As in our baseline specification, accounting for location selection drops the estimate up to 50%, when the correction functions are approximated by a fourth order polynomial. The selection-corrected estimate of the local bias coefficient (δ^{dev}) is about -0.22 bps, so that the economic effect of a one standard deviation increase in distance is about -2.5 bps ($\approx -0.22 \times 11.45$). That is, the local bias estimate of the above portfolio choice

is -0.00702 when the location choice model does not include either income or job codes. Both are slightly lower in magnitude than the distance coefficient obtained from the baseline location choice model, which is -0.00731.

model is about three times higher than the local bias estimate of the Tobit.

We also extend our analysis to Russell 3000 stocks, 80 MSAs and extended household demographics and find similar results in Appendix Table 2.5.

Appendix Table 2.1: Tobit Estimation of Household Portfolio Choice with Full Controls and Distance Indicator Variables

This table presents the estimation results from four Tobit models of household portfolio choice with controls for household demographics, stock characteristics and MSA demographics. In Columns 1 and 3, the specifications do not correct for location selection. In Columns 2 and 3, the specifications correct for location selection with a quartic approximation of the correction functions. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Away, an indicator variable that equals one if the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code is greater than a specific threshold value. In columns (1) to (2), the distance (away) threshold is 100 miles, and in columns (3) to (4) it is 250 miles. The household demographics controls are the log of a household's income (LogIncome), the log of the age of a household head (LogAge), fixed effects for the occupation of a household's head (Managerial to BlueCollar), gender (Male) and marital status (Married), and the number of kids under the age of 18 a household has (Kids). The stock characteristics controls are: the log of the price (LogPrice), the log of the market capitalization (LogSize), the book-to-market ratio (BTM), the turnover ratio (Turnover), the momentum of the past 12 months (Momentum), the volatility (Volatility) and industry fixed effects based on the 17 Fama-French industry portfolios. Controls for MSA demographics include: the log of the population number (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), the log of the housing price index (LogHPI), where the subscripts c and h denote respectively the MSAs of household i's residence and stock j's headquarters. The constant term and the standard deviation of the normal error term in the Tobit regression are not reported for brevity. The sample period is January 1991 to November 1996.

	100 Miles Away		250 Miles Away	
	(1)	(2)	(3)	(4)
Away	-0.629	-0.429	-0.478	-0.266
	[-7.1]	[-8.77]	[-7]	[-6.95]
LogIncome	0.027	0.02	0.026	0.021
	[2.4]	[1.73]	[2.39]	[1.86]
LogAge	0.106	0.109	0.099	0.107
	[5.08]	[5.27]	[4.94]	[5.17]
Male	0.033	0.033	0.029	0.029
	[2.33]	[2.58]	[1.97]	[2.33]
Married	-0.012	-0.005	-0.014	-0.007
	[-1.11]	[-0.49]	[-1.25]	[-0.69]
Kids	-0.01	-0.011	-0.012	-0.012
	[-1.78]	[-1.95]	[-2.02]	[-2.14]
LogPrice	-0.38	-0.374	-0.381	-0.374
	[-20.29]	[-23.43]	[-20.39]	[-23.16]
LogSize	0.428	0.426	0.428	0.426
	[40.12]	[47.47]	[38.66]	[45.98]
BTM	0.099	0.101	0.098	0.101
	[4.98]	[5.47]	[4.73]	[5.26]
Turnover	0.356	0.348	0.364	0.351
	[19.55]	[22.52]	[19.51]	[22.89]
Momentum	-0.034	-0.035	-0.035	-0.035
	[-1.31]	[-1.64]	[-1.28]	[-1.58]
Volatility	-0.003	-0.005	-0.008	-0.006
	[0.12]	[0.43]	[0.18]	[0.14]
LogPop_h	-0.005	-0.023	-0.001	-0.026
	[-0.35]	[-1.21]	[-0.1]	[-1.29]
LogUnemp_h	-0.016	-0.011	-0.011	-0.008
	[-1.9]	[-1.48]	[-1.48]	[-1.14]
Incomepc_h	-0.012	-0.01	-0.009	-0.008
	[-3.35]	[-23.73]	[-2.84]	[-15.85]
LogHPI_h	-0.183	0.015	-0.414	-0.075
	[-0.45]	[-0.06]	[-0.9]	[-0.36]
LogPop_c	-0.021	-0.094	-0.026	-0.111
	[-1.44]	[-4.16]	[-1.72]	[-5.31]
Unemp_c	-0.004	0.015	-0.004	0.017
	[-0.35]	[1.43]	[-0.42]	[1.95]
Incomepc_c	-0.008	-0.01	-0.005	-0.01
	[-1.9]	[-11.29]	[-1.21]	[-13.5]
LogHPI_c	-0.44	0.066	-0.773	-0.047
	[-0.15]	[0.25]	[-0.49]	[7.49]
HH Occupation F.E.	YES	YES	YES	YES
Stock Industry F.E.	YES	YES	YES	YES

Appendix Table 2.2: Conditional Logit Estimation of Household Location Choice, Alternative Location-Choice Models

This table presents the results from the maximum likelihood estimation of three alternative conditional logit models for household location choices. t-statistics [shown in brackets] are based on standard errors clustered at the household level. The choice set of households in Columns 1 and 2 consists of 57 MSAs, while in Column 3 of 53 MSAs. The dependent variable is an indicator variable that equals one if a household resides in the specific MSA. In Column 1, we exclude household income (and its interactions with MSA demographics) from the explanatory variables. In Column 2, we exclude household income and job codes (and their interactions with MSA demographics) from the explanatory variables. All other explanatory variables are the same as in Table 2.3. Log-likelihood is the estimated log-likelihood value. Pseudo R^2 is the McFadden's pseudo R^2 measure based on the estimated log-likelihood value. I is the number of households.

	(1)	(2)
LogPop	0.323	0.349
	[1.33]	[1.44]
Unemp	-0.512	-0.514
	[-3.03]	[-3.04]
Incomepc	0.322	0.316
	[5.78]	[5.70]
LogHPI	6.831	6.857
	[3.41]	[3.43]
LogPop×LogAge	0.147	0.155
	[2.35]	[2.47]
LogPop×Managerial	0.099	
	[3.18]	
LogPop×SalesSvc	0.035	
	[0.70]	
LogPop×WhiteCollar	0.195	
	[2.99]	
LogPop×BlueCollar	0.069	
	[0.96]	
LogPop×Male	0.061	0.048
	[1.25]	[1.00]
LogPop×Married	-0.046	-0.047
	[-1.37]	[-1.39]
LogPop×Kids	-0.012	-0.013
	[-0.78]	[-0.83]
Unemp×LogAge	0.095	0.099
	[2.21]	[2.29]
Unemp×Managerial	0.002	
	[0.11]	
Unemp×SalesSvc	-0.005	
	[-0.15]	
Unemp×WhiteCollar	0.060	
	[1.49]	
Unemp×BlueCollar	0.121	
	[3.17]	
Unemp×Male	0.012	0.011
	[0.32]	[0.30]
Unemp×Married	-0.012	-0.015
	[-0.54]	[-0.67]
Unemp×Kids	0.018	0.017
	[1.72]	[1.65]

Appendix Table 2.2 Cont'd: Conditional Logit Estimation of Household Location Choice, Alternative Location-Choice Models

	(1)	(2)
Incomepc×LogAge	-0.048	-0.048
	[-3.27]	[-3.27]
Incomepc×Managerial	-0.006	
	[-0.83]	
Incomepc×SalesSvc	-0.021	
	[-1.68]	
Incomepc×WhiteCollar	-0.016	
	[-0.98]	
Incomepc×BlueCollar	-0.02	
	[-1.13]	
Incomepc×Male	-0.047	-0.046
	[-4.51]	[-4.42]
Incomepc×Married	-0.022	-0.022
	[-2.80]	[-2.75]
Incomepc×Kids	0.001	0.001
	[0.22]	[0.26]
LogHPI×LogAge	-1.281	-1.321
	[-2.50]	[-2.58]
LogHPI×Managerial	-0.199	
	[-0.81]	
LogHPI×SalesSvc	-0.391	
	[-1.01]	
LogHPI×WhiteCollar	-0.026	
	[-0.05]	
LogHPI×BlueCollar	-1.691	
	[-3.03]	
LogHPI×Male	-0.694	-0.744
	[-1.50]	[-1.61]
LogHPI×Married	-0.308	-0.285
	[-1.06]	[-0.99]
LogHPI×Kids	-0.509	-0.508
	[-4.14]	[-4.14]
Log-likelihood	-38,650.4	-38,679.2
Pseudo R^2	0.108	0.107
I	10,712	10,712

Appendix Table 2.3: Tobit Estimation of Household Portfolio Choice with Correction for Location Selection, Full Controls, No Income or No Income and Job Code in the Location Model

This table presents the estimation results from six Tobit models of household portfolio choice that correct for location selection based on a location choice model that excludes either the income or the income and job code of a household. To derive the predicted location probabilities of the households, we exclude either the income (Columns 1, 2 and 3) or the income and occupation codes (Columns 4, 5 and 6) of the households as explanatory variables in the first-stage conditional logit estimation. The correction functions are approximated linearly (Columns 1 and 4), quadratically (Columns 2 and 5) and quartically (Columns 3 and 6). The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. Household demographics, stock characteristics and MSA demographics are included as controls. The household demographics controls LogIncome to Kids, the stock characteristics controls LogPrice to Volatility, the industry controls, and the controls for MSA characteristics LogPop_c to LogHPI_h are the same as those shown in Table 2.4. The constant term and the standard deviation of the normal error term in the Tobit regression are not reported for brevity. The sample period is January 1991 to November 1996.

	Dependent Variable: Portfolio Weight $w_{i,j}$					
	No Income			No Income and Job Code		
	(1) Linear Ψ's	(2) Quadratic Ψ's	(3) Quartic Ψ's	(4) Linear Ψ's	(5) Quadratic Ψ's	(6) Quartic Ψ's
Distance	-0.012	-0.009	-0.007	-0.012	-0.009	-0.007
	[-6.33]	[-6.9]	[-9.19]	[-6.31]	[-7.22]	[-10.39]
LogIncome	0.028	0.027	0.023	0.028	0.027	0.024
	[2.66]	[2.55]	[2.3]	[2.68]	[2.55]	[2.42]
LogAge	0.102	0.103	0.107	0.101	0.103	0.107
	[4.87]	[4.83]	[5.34]	[4.83]	[4.97]	[5.35]
Male	0.019	0.033	0.032	0.02	0.031	0.03
	[1.26]	[2.26]	[2.48]	[1.37]	[2.23]	[2.47]
Married	-0.026	-0.01	-0.007	-0.025	-0.012	-0.009
	[-2.21]	[-0.83]	[-0.65]	[-2.15]	[-1.05]	[-0.91]
Kids	-0.015	-0.013	-0.012	-0.015	-0.013	-0.012
	[-2.46]	[-2.15]	[-2.11]	[-2.43]	[-2.17]	[-2.25]
LogPrice	-0.378	-0.376	-0.373	-0.378	-0.377	-0.373
	[-19.36]	[-20.05]	[-25.76]	[-19.4]	[-20.14]	[-26.01]
LogSize	0.429	0.428	0.424	0.429	0.428	0.425
	[38.58]	[41.11]	[47.73]	[38.56]	[41.23]	[50.06]
BTM	0.102	0.101	0.103	0.102	0.101	0.102
	[4.83]	[4.78]	[5.53]	[4.83]	[4.8]	[5.69]
Turnover	0.369	0.366	0.351	0.369	0.367	0.354
	[20.71]	[23.41]	[23.73]	[20.76]	[23.22]	[24.31]
Momentum	-0.037	-0.036	-0.036	-0.037	-0.036	-0.036
	[-1.3]	[-1.31]	[-1.82]	[-1.31]	[-1.29]	[-1.66]
Volatility	0.005	0.008	-0.004	0.005	0.007	-0.001
	[0.2]	[0.28]	[0.27]	[0.2]	[0.27]	[0.3]
LogPop_c	0.015	-0.012	0.002	0.091	0.084	0.083
	[0.8]	[-0.51]	[0.05]	[1.48]	[1.41]	[1.73]
Unemp_c	-0.002	-0.003	-0.006	-0.144	-0.145	-0.157
	[-0.2]	[-0.53]	[-0.87]	[-3.65]	[-3.53]	[-4.22]
Incomepc_c	0.003	0.001	0	0.193	0.19	0.21
	[1.02]	[-0.11]	[-0.17]	[3.05]	[3.2]	[3.92]
LogHPI_c	-0.278	-0.117	-0.033	0.117	0.127	0.136
	[0.53]	[0.75]	[0.27]	[2.36]	[2.73]	[3.36]
LogPop_h	0.025	0.017	-0.179	-0.099	-0.089	-0.078
	[0.68]	[0.01]	[-6.55]	[-1.86]	[-1.66]	[-1.45]
Unemp_h	0.011	0.009	0.029	-0.092	-0.099	-0.089
	[0.89]	[1.05]	[3.53]	[-2.55]	[-2.99]	[-3.12]
Incomepc_h	0.008	0.006	-0.011	0.074	0.073	0.084
	[2.25]	[1.47]	[-19.68]	[1.17]	[1.2]	[1.68]
LogHPI_h	-0.493	-0.269	-0.04	-0.243	-0.237	-0.235
	[0.17]	[-0.14]	[0.47]	[-4.39]	[-4.33]	[-4.57]
HH Occupation F.E.	YES	YES	YES	YES	YES	YES
Stock Industry F.E.	YES	YES	YES	YES	YES	YES

Appendix Table 2.4: Regressions of Household Portfolio Weight Deviations from the Market

This table presents the estimation results from four linear regressions of household portfolio weight deviations from the market, with full controls and correcting for location selection. The correction functions are approximated with polynomials. In Columns 1, 4 and 7, the approximation of the correction functions is linear. In Columns 2, 6 and 8, the approximation of the correction functions is quadratic. In Columns 3, 7 and 9, the approximation of the correction functions is quartic. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j} - w_j^{VW}$, i.e. the difference of the value weighted portfolio weight of the market on stock j from the the portfolio weight of household i residing in city c on the same stock. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. Other variables include controls for household demographics, stock characteristics, and MSA demographics. The household demographics controls are: the log of a household's income (LogIncome), the log of the age of a household head (LogAge), fixed effects for the occupation of a household's head (Managerial to BlueCollar), gender (Male) and marital status (Married), and the number of kids under the age of 18 a household has (Kids). The stock characteristics controls are: the log of the price (LogPrice), the log of the market capitalization (LogSize), the book-to-market ratio (BTM), the turnover ratio (Turnover), the momentum of the past 12 months (Momentum), the volatility (Volatility) and industry fixed effects based on the 17 Fama-French industry portfolios. Controls for MSA demographics include: the log of the population number (LogPop), the unemployment rate (Unemp), the income per capita (Incomepc), the log of the housing price index (LogHPI), where the subscripts c and h denote respectively the MSAs of household i's residence and stock j's headquarters. The estimated values of the coefficients in the correction functions and the constant in the linear regressions are not reported for brevity. The sample period is January 1991 to November 1996.

	Dependent Variable: $w_{i,j} - w_j^{VW}$ (Coefficient Estimates in bps)			
	(1) Uncorrected	(2) Linear Ψ's	(3) Quadratic Ψ's	(4) Quartic Ψ's
Distance	-0.429	-0.354	-0.289	-0.217
	[-6.33]	[-5.96]	[-6.7]	[-7.74]
LogIncome	-0.181	0.394	0.007	-0.295
	[-1.17]	[1.79]	[0.04]	[-1.43]
LogAge	-1.025	-0.792	-0.826	-0.726
	[-3.16]	[-2.29]	[-2.55]	[-2.24]
Male	-0.654	-0.907	-0.665	-0.581
	[-2.48]	[-3.43]	[-2.4]	[-2.29]
Married	0.057	-0.495	-0.02	0.249
	[0.19]	[-2.39]	[-0.15]	[0.92]
Kids	0.027	-0.025	0.034	0.042
	[0.33]	[-0.3]	[0.45]	[0.52]
LogPrice	-1.734	-1.638	-1.635	-1.53
	[-3.46]	[-3.23]	[-3.12]	[-2.82]
LogSize	-5.82	-5.818	-5.825	-5.836
	[-15.89]	[-16.18]	[-16.76]	[-16.97]
BTM	0.806	0.786	0.768	0.767
	[6.05]	[5.61]	[5.57]	[5.36]
Turnover	10.6	10.355	10.238	9.858
	[11.72]	[11.72]	[12.9]	[12.75]
Momentum	-1.536	-1.579	-1.563	-1.582
	[-3.17]	[-3.24]	[-3.17]	[-3.14]
Volatility	1.127	1.112	1.167	1.229
	[0.63]	[0.63]	[0.66]	[0.62]
LogPop_c	-1.2	-0.119	-0.172	-0.588
	[-2]	[-0.18]	[-0.24]	[-0.76]
Unemp_c	0.377	0.275	0.13	0.143
	[1.12]	[1.01]	[0.58]	[0.55]
Incomepc_c	0.179	0.206	0.194	0.133
	[1.15]	[1.42]	[1.54]	[1.43]
LogHPI_c	-11.397	-9.432	-10.494	-4.941
	[1.07]	[1.21]	[1.45]	[1.05]
LogPop_h	-1.517	0.333	0.756	-3.014
	[-2.09]	[0.13]	[0.31]	[-2.38]
Unemp_h	0.813	0.775	0.588	1.106
	[1.83]	[1.67]	[1.39]	[2.82]
Incomepc_h	0.236	0.354	0.347	0.064
	[1.29]	[2.1]	[1.94]	[0.38]
LogHPI_h	-35.526	-31.756	-31.079	-9.755
	[-1.23]	[-1.03]	[-0.89]	[-0.48]
HH Occupation F.E.	YES	YES	YES	YES
Stock Industry F.E.	YES	YES	YES	YES

Appendix Table 2.5: Tobit Estimation of Household Portfolio Choice, without and with Correction for Location Selection, for Russell 3000, Households Residing in 80 MSAs and Extended Household Demographic Controls

This table presents the estimation results from eight Tobit models of household portfolio choice, without and with location selection, for stocks in Russell 3000 and extended household demographic controls. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on standard errors clustered at the household and MSA level. The dependent variable is $w_{i,c,j}$, the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address ZIP Code. In Columns 1, 3, 5 and 7, there is no correction for location selection. In Columns 2, 4, 6 and 8, we correct for location selection, approximating the correction functions with fourth order polynomials. In Columns 1 and 2, there are no controls. In Columns 3 and 4, household demographics are included as controls. In Columns 5 and 6, stock characteristics are added as controls. In Columns 7 and 9, MSA demographics are added as controls. The household demographic controls are the same as before. They additionally contain the percentage of the household's ZIP Code population that holds a bachelor's or higher degree (Educ), fixed effects for the percentages of the household's ZIP Code population that are Black (Black), Hispanic (Hispanic) or Asian or other race (AsianOther) (with the base group being the percentage of the population that is White), and the professional industrial proximity to stock j of the household's ZIP Code population (ProfProxim). The stock characteristics and the MSA demographics are the same as before. The estimated values of the coefficients in the employed correction functions, the constant and the standard deviation of the normal error in the Tobit regressions are not reported for brevity. Panel B presents the average marginal effects on $w_{i,c,j}$ in basis points and as percentage of the mean. The sample period is January 1991 to November 1996. The universe of stocks is comprised by the publicly traded firms that were *ever* included in the Russell 3000 index during the sample period. Households reside in 80 MSAs with a population of at least 500K at the beginning of the year 1991.

	Panel A: Coefficient Estimates							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	No Controls		Plus HH Demo's		Plus Stock Char's		Plus MSA Demo's	
Distance	-0.011	-0.007	-0.012	-0.007	-0.012	-0.007	-0.012	-0.007
	[-5.92]	[-6.37]	[-5.64]	[-8.8]	[-5.41]	[-7.31]	[-5.57]	[-7.03]
Educ			0.11	0.102	0.095	0.082	0.099	0.079
			[2.32]	[2.97]	[2.08]	[2.45]	[2.55]	[2.43]
ProfProxim			0.486	0.293	0.34	0.083	0.314	0.064
			[3.78]	[3.73]	[3.92]	[1.24]	[3.87]	[0.98]
HH Demographics	NO	NO	YES	YES	YES	YES	YES	YES
HH Occupation F.E.	NO	NO	YES	YES	YES	YES	YES	YES
ZIP Code Race F.E.	NO	NO	YES	YES	YES	YES	YES	YES
Stock Characteristics	NO	NO	YES	YES	YES	YES	YES	YES
MSA Demographics	NO	NO	YES	YES	YES	YES	YES	YES
Corr. Func.	NO	YES	NO	YES	NO	YES	NO	YES
	Panel B: Average Marginal Effects in Basis Points and as Percentage of the Mean							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	No Controls		Plus HH Demo's		Plus Stock Char's		Plus MSA Demo's	
Distance	-1	-0.57	-1.04	-0.59	-0.42	-0.24	-0.44	-0.25
	-34.9%	-19.9%	-36.3%	-20.6%	-14.7%	-8.4%	-15.4%	-8.7%

Chapter 3

Estimates of the Costs and Benefits of Information in Stock Markets*

3.1. Introduction

The asset management industry (or *buy-side*) is one of the largest in the world. According to a Boston Consulting Group survey featured in the *Financial Times*, in 2015 it generated profits of 102 billion dollars globally, from assets under management of 71.4 trillion dollars.[1] A large part of the industry are mutual funds that manage household wealth.[2] Even after the 2008 Financial Crisis, the fraction of households worldwide investing in mutual funds continued to rise. For instance, the Investment Company Institute (ICI) reports that 54.9 million U.S. households (or 43.6%) owned mutual funds in 2016.[3] About 70% of these assets are managed actively, whereby

*Co-authored with Harrison Hong, John R. Eckel Jr. Professor of Financial Economics in the Department of Economics at Columbia University, and Jiangmin Xu, Assistant Professor of Finance at the Guanghua School of Management of Peking University.

We thank Motohiro Yogo, Peter Kondor, Russ Wermers, Jakub Kastl and participants at the Finance and Industrial Organization PhD Workshops at Princeton University and the 12th Annual Hedge Fund Conference.

[1] See "Global fund industry manages record $74tn", *Financial Times*, July 7, 2015.

[2] For the US alone in 2007, management fees were equal to 1 percent of 12 trillion dollar assets (Greenwood and Scharfstein (2013)).

[3] ICI Research Perspective, October 2016, available at www.ici.org/pdf/per22-06.pdf

managers attempt to pick winning stocks based on research.[4] As a result, active equity funds hold the largest stake of corporate equity and play a pivotal role in the determination of stock prices (e.g., Grinblatt, Titman, and Wermers (1995), Wermers (1999), Gompers and Metrick (2001)).

In this paper, we tackle a question that is fundamental to understanding the role of these institutional investors more generally, and mutual funds more specifically in the economy: *How do money managers generate their research ideas about what stocks to buy?* Interestingly, much of this research is not created by money managers themselves, but is bought from security analysts working for investment banks (i.e. the *sell-side*). According to a poll by Frost Consulting and Quark featured in *Financial News*, 38 out of 50 European chief investment officers [of mutual funds] said that "more than 80% of their research came from investment banks" in 2013.[5]

Such polls are consistent with the industry estimates of aggregate expenditure on equity research. In the same *Financial News* article, Frost Consulting and Quark estimated that mutual funds globally spend annually about 22 billion dollars for research services, or roughly 20% of profits. Overall, according to estimates of Edison Research Investment presented in the same article, 9,000 analysts were employed in that year worldwide. The 22 billion dollar estimate of equity research spending is only a lower bound of the total money going to the sell-side. This is because managers also pay commissions to brokerage houses for trade executions based on sell-side research.

Yet, the literature on institutional investor portfolios has neglected the importance of sell-side research for buy-side portfolios (e.g., Grinblatt and Titman (1989), Mankiel (1995), Falkenstein (1996), Daniel, Grinblatt, Titman, and Wermers (1997), French (2008), Fama and French (2010), Shumway, Szefler, and Yuan (2011)). This is likely due to a lack of data on individual fund expenditure on equity research. While there

[4]Morningstar Direct^{SM} Asset Flows Commentary: United States, December 2016, available at http://corporate.morningstar.com/US/documents/AssetFlows/AssetFlowsDec2016.pdf

[5]See "FCA's commissions review puts spotlight on sellside", *(London) Financial News*, November 6, 2013.

is no direct data on such spending, we show in this paper that we can, nonetheless, measure the importance of the sell-side by modeling and estimating the effect of analyst coverage of stocks on buy-side portfolios.

Our model has two stages through which the fund portfolio is determined. The first stage features sell-side analysts pitching stocks to managers as investment ideas. The premise is that there are fixed costs to investing in stocks. For instance, in our data, the universe has 4,208 available stocks for consideration, and yet a typical active equity mutual fund owns only 89. Even if one accounts for benchmarks that might limit the universe of stocks for a manager, there is an extreme degree of sparsity in institutional investor portfolios.

We model the extensive margin of manager portfolios and the purchase of sell-side research or investment ideas in affecting this margin. The direct marketing process of covered stocks determines the choice set from which managers will create their portfolios. The outcome of stock pitching is assumed to be probabilistic. Managers have limited time and thus cannot talk to all the numerous analysts or read all the research about all the numerous stocks the latter cover. We model the probability of having a stock in the manager's consideration set as a logistic function of analyst coverage. Our first-stage choice set formation is analogous to the set-up in Goeree (2008), who models the effect of advertising expenditure on the probability that a product ends up in a household's consumer basket.[6] The key idea is that the higher the coverage of a stock, the more likely that this stock gets pitched to a manager and to draw their attention.[7]

The second part is that conditioning on meetings with analysts, which determine

[6]See Kastl (2016) for a recent literature review on the intersection of Industrial Organization with Finance.

[7]Our modeling of the extensive margin can also be motivated by recent research on limited or rational inattention (see, e.g., Sims (2003), Peng and Xiong (2006), Gabaix (2016)) and particularly to fund managers (Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016)). Our model emphasizes then the supply or sell-side dimension whereby analyst coverage can remedy these attention constraints.

the consideration set of stocks for a manager, risk-neutral managers then form portfolios in the face of quadratic trading costs and short-selling constraints.[8] That is, once a stock is pitched to a manager, it is on his radar and he then determines his portfolio weight on the stock. Here, we also allow analyst coverage to affect the portfolio weights of stocks in the consideration set, i.e. the intensive margin, by improving the precision of the private signals of the manager.

Our model of the intensive margin is similar to contemporaneous work by Hong and Xu (2015), who model the effect of social networks on the precision of managerial signals in the standard linear-normal signal extraction framework.[9] The signal might end up being positive or negative, but the expected portfolio weight of a stock increases in analyst coverage. The reason is censoring due to short-sales constraints: as a manager's precision of private signal increases, he places more weight on these signals relative to his prior. Since there are shorting constraints, we only see the positive latent demand from the increased precision and hence the higher expected portfolio weight.

In short, sell-side coverage of a stock might then influence managerial portfolios in two distinct ways. First, coverage makes it more likely that a stock is included in the consideration set of a manager by drawing managerial attention to the stock (the extensive margin). Second, coverage leads to more precise private signals (the intensive margin). Our model yields a tractable logistic-mixed Tobit regression with heteroskedasticity (i.e., a specification in the class of mixed censored models with heteroskedasticity) of managerial portfolio weights.

Estimating this model requires that we address the endogeneity of analyst cover-

[8]Both of these assumptions are realistic for fund managers. They are prohibited from shorting by institutional charter (see, e.g., Chen, Hong, and Stein (2002) and large risk positions have to incur price impact or hit institutional risk management constraints (Gârleanu and Pedersen (2013)).

[9]Similar to Brandt, Santa-Clara, and Valkanov (2009), Hjalmarsson and Manchev (2012), Gârleanu and Pedersen (2013), Hortaçsu and Kastl (2015), and Koijen and Yogo (2016), our model expresses mutual fund portfolio weights as a function of expected return factors and an idiosyncratic latent demand shock that represents the manager's private information.

age. We thus use as an instrument the mergers of brokerage firms from Hong and Kacperczyk (2010). Treated stocks typically experience a reduction of coverage by one analyst as the merged broker streamlines the workforce and fires one of the two analysts covering an industry (e.g. oil or technology). We follow a control function approach in the spirit of Petrin and Train (2010). In particular, our identification strategy is to express the correlations between analyst coverage and the latent factors in the demand for stocks as functions of residuals from a quasi-experiment. We then plug these residuals in our demand model as additional covariates.

We first consider a difference-in-difference regression to see how aggregate mutual fund ownership and stock price rely on an exogenous shock to analyst coverage. A merger of brokerage houses results in a decrease of: (i) analyst coverage by 0.65 analysts, (ii) aggregate mutual fund ownership (MFO) of a stock by 1.34%, and (iii) price by 17%. Given that MFO is on average 5% and the average price is approximately 20 dollars (see summary statistics in Table 3.1), the magnitude of these reduced form results suggest that analyst coverage is an important determinant of the aggregate equity demand in the mutual fund industry (i.e, firing 1 analyst reduces a stock's MFO by approximately 2.06% and price by 26.2%).

We then show using our structural model of a fund portfolio that both channels (fixed and variable cost of information) are statistically significant based on our nested likelihood ratio tests. Our tractable model allows for a decomposition of the effect of analyst coverage on the extensive and intensive margins of ownership. The economic effect on the extensive margin is larger, accounting for more than 60% (and up to 80%) of the total on average (about 0.74 bps when the average portfolio weight on a stock is 3.2 bps) depending on the controls.

Our model naturally predicts that the performance of fund holdings is increasing in analyst coverage. The reason is that managers are more likely to have spent money to obtain the information from analysts on these stocks and hence their performance

should be higher as a result. To this end, we conduct a fund holdings performance test where we compare the within-manager fund holdings performance using the methodology in Coval and Moskowitz (2001) and Cohen, Frazzini, and Malloy (2008). Among stocks held by a manager, those with greater analyst coverage outperform those with lower coverage on average by 39 bps per month (or roughly 5% per annum alpha).[10]

Our paper contributes to the literature in two distinct ways. First, we bring attention to the importance of sell-side research for mutual fund portfolios. Second, and in doing so, we develop a tractable structural model of mutual fund portfolio weights that calls attention to the need to carefully model the extensive and intensive margins of extremely sparse mutual fund portfolios.

3.2. Model

3.2.1. Consideration Set Formation

In every period, manager i considers investing in only a subset of stocks $\mathcal{J}_i \subset \mathcal{J}$. Analysts affect this choice set through a direct marketing process of the stocks they cover. However, even if stock $j \in \mathcal{J}$ is covered by a number of analysts A_j, it is not taken for granted that manager i will consider it; in reality, there are thousands of analysts promoting thousands of stocks, while managers have only limited time.

To capture this element, we introduce a random idiosyncratic consideration factor, $\epsilon_{i,j}$, which makes manager i' attention on stock j stochastic. We denote the outcome

[10] Our return findings complement but are distinct from those in Lee and So (2017), who show that the stocks with greater abnormal analyst coverage as in Hong, Lim, and Stein (2000) out-perform stocks with abnormally low coverage. What matters for the marketing and precision channels in our model is the number of analysts covering a stock and not the residual coverage. These recent results overturn the neglected firm effect of Arbel, Carvell, and Strebel (1983), who found using an earlier sample that stocks with zero or low analyst coverage out-perform stocks with high analyst coverage. But using analysts coverage as a proxy for neglect is potentially problematic since fund managers pay for coverage. The neglected firm effect more likely is connected to media coverage, which are consumed more broadly by retail investors and has been shown to be associated with over-pricing and lower expected returns by Fang, Peress, and Zheng (2014). They also document that manager holdings of high media coverage stocks underperform in contrast to our findings.

of this process by $d_{i,j} \in \{0,1\}$. If $d_{i,j} = 1$, stock j is successfully pitched to manager i as an investment idea. On the other hand, $d_{i,j} = 0$ immediately implies a zero portfolio weight on the stock. The exact consideration process of manager i for stock j is assumed to be:

$$d_{i,j} = \mathbf{1}\left[\kappa_i + \lambda_i A_j + \epsilon_{i,j} \geq 0\right] \tag{3.1}$$

where $\mathbf{1}[\cdot]$ is an indicator function and $\boldsymbol{\theta}_i^{cons} \equiv (\kappa_i, \lambda_i)$ are manager i's asset consideration parameters. In our baseline model, we only let analyst coverage influence the extensive margin of buy-side portfolios. In a Robustness Section, we also allow other stock characteristics to be potentially influential as well.

In principle, we expect the impact of analysts on the probability of considering a stock to be positive, i.e. it is more likely that a manager considers a stock for investment if there are more analysts covering that stock. However, managers are heterogeneous, so the direct marketing effect of coverage for each of them is in general different. Under the assumption that, conditional on the number of analysts A_j, $\epsilon_{i,j}$ is independently and identically distributed according to the standard logistic distribution, the probability with which manager i includes stock j in his consideration set, $\mathcal{J}_i$, takes the following logistic form:[11]

$$\pi_{i,j} = \frac{\exp\left(\kappa_i + \lambda_i A_j\right)}{1 + \exp\left(\kappa_i + \lambda_i A_j\right)} \tag{3.2}$$

In that way, the technology that determines the choice set formation process of a manager is in the spirit of Goeree (2008). Nevertheless, the number of analysts who cover stock j is endogenous and thus possibly correlated with the latent consideration factor of manager i for that stock, $\epsilon_{i,j}$. To this end, we also use an instrument for A_j, when we run the estimation of the model. Specifically, in line with Hong and Kacperczyk (2010), we consider *mergers of brokerage houses* as a source of exogenous

[11] We assume away any possible attention spillovers between stocks.

variation for analyst coverage.

3.2.2. Portfolio Choice

Once manager i pays attention to a set of stocks, $\mathcal{J}_i$, he makes his investment decisions. We assume that all managers are *risk neutral* and have *subjective* beliefs about the one-period ahead returns of stocks. We also assume the existence of a risk-free asset.

In particular, absent analyst coverage, manager i has a normal prior for stock j's excess return, $\tilde{f}_j \equiv \tilde{R}_j - R_f$, based on stock j's financial characteristics, $\boldsymbol{X}_j$, and his precision, τ_i^0:

$$\tilde{f}_j \,|\boldsymbol{X}_j \sim \mathcal{N}\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j, \frac{1}{\tau_i^0}\right) \tag{3.3}$$

However, thanks to the analysts, manager i receives a private signal for stock j's return:

$$S_{i,j} = \tilde{f}_j + \eta_{i,j} \tag{3.4}$$

Conditional on stock j's risk factors $\boldsymbol{X}_j$ and analyst coverage A_j, the signal's noise, $\eta_{i,j}$, is assumed to be normally distributed with a zero mean and a precision that is expected to increase with the number of the stock's analysts , $\tau_i^{\eta}(A_j)$. Consequently, by the projection theorem, manager i's updated subjective expectation for stock j's excess return is:

$$\mathbb{E}\left(\tilde{f}_j \,|\boldsymbol{X}_j, A_j, S_{i,j}\right) = \alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j} \tag{3.5}$$

where, since the private signal, $S_{i,j}$, is unobservable to the econometrician, we define manager i's *latent demand* for stock j as:

$$\xi_{i,j} \equiv \frac{\tau_i^{\eta}(A_j)}{\tau_i^0 + \tau_i^{\eta}(A_j)} \left[S_{i,j} - (a + \boldsymbol{b}\boldsymbol{X}_j) \right] \tag{3.6}$$

The latent demand, $\xi_{i,j}$, reflects manager i's private information about stock j as well as his optimism or pessimism about the stock's prospects. Yet, here, Equations (3.3) and (3.4) notably imply that:

$$\xi_{i,j} \left| \boldsymbol{X}_j, A_j \right. \sim \mathcal{N} \left(0, \underbrace{\frac{\tau_i^{\eta}(A_j)}{\tau_i^0 \left[\tau_i^0 + \tau_i^{\eta}(A_j) \right]}}_{\sigma_{i,j}^2} \right) \tag{3.7}$$

so that the variance of manager i's latent demand for stock j (namely, $\sigma_{i,j}^2$) is expected to increase with stock j's analyst coverage (given that the precision of the signal is conjectured to increase as well).[12]

By the risk neutrality, Equation (3.5) is the only relevant moment of stock j's return for manager i's objective, which is to maximize the expected excess return of his fund's assets under the presence of short-selling constraints and trading costs. Indeed, mutual funds cannot short by institutional charter, so that manager i's portfolio weight on a stock cannot be negative, i.e. $\{w_{i,j} \geq 0\}_{j \in J_i}$ (Chen, Hong, and Stein (2002)). Moreover, in the spirit of Gârleanu and Pedersen (2013), we assume that managers face quadratic costs from transactions. For our case, we specify the quadradicity of transaction costs in terms of acquired market value rather than number of shares. That is, when manager i acquires $n_{i,j}$ shares of stock j, he pays:

$$TC_{i,j} = \frac{1}{2} \Lambda_i \left(P_j n_{i,j} \right)^2 \tag{3.8}$$

where Λ_i scales the level of transaction cost that manager i faces for every unit of money spent on stock j. Conveniently, we let that scaling factor be:

[12] That is, $\frac{d\sigma_{i,j}^2(A_j)}{dA_j} = \frac{\tau_i^{\eta}(A_j)'}{\left[\tau_i^0 + \tau_i^{\eta}(A_j)\right]^2} > 0$, if $\tau_i^{\eta}(A_j)' > 0$.

$$\Lambda_i = \frac{c_i}{W_i^2} \tag{3.9}$$

The more assets under management a fund can attract, the less it has to worry about transaction costs. If the dependence on manager i's total assets, W_i, is quadratic, transaction costs matter much less for richer funds and essentially refer to the level of portfolio weights.

Combining Equations (3.5), (3.8) and (3.9), we end up expressing manager i's objective as follows:[13]

$$\max_{\{w_{i,j}\geq 0\}_{j\in\mathcal{J}_i}} \left\{ \sum_{j\in\mathcal{J}_i} \left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}\right) w_{i,j} - \frac{1}{2} c_i w_{i,j}^2 \right] \right\} \tag{3.10}$$

The KKT conditions then imply that:

$$w_{i,j} = \left(\frac{\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}}{c_i} \right)^+ \tag{3.11}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ simultaneously captures manager i's decision of whether and how much to invest in every stock included in his consideration set. According to Equation (3.11), manager i invests in stock j provided that the expected excess return of the stock is positive in his view. The transaction cost to which he is subject determines the actual level of the portfolio weight, scaling it up or down.

Since from Equation (3.7) the latent demand, $\xi_{i,j}$, is normally distributed, Equation (3.11) constitutes a heteroskedastic tobit model, mixed with the first-stage logit model that determines the consideration set. However, as the model stands, it is under-identified. To this end, we make the following simplifications; First, we normalize manager i's precision, τ_i^0 and transaction cost parameter, c_i, to be equal to 1. And second, we parameterize the precision of the signal's noise, $\eta_{i,j}$, as follows:

[13] Manager i's budget constraint requires the sum of his stock-portfolio weights to equal one minus the weight on the risk-free asset. The latter is assumed to be perfectly adjustable to the needs of the portfolio optimization problem.

$$\tau_i^{\eta}(A_j) = \frac{\exp[2\left(\gamma_i + \delta_i A_j\right)]}{1 - \exp[2\left(\gamma_i + \delta_i A_j\right)]} \tag{3.12}$$

Equations (3.7) together with (3.11) entails that:

$$w_{i,j} = \left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \exp\left(\gamma_i + \delta_i A_j\right) \zeta_{i,j}\right)^+ \tag{3.13}$$

where $\zeta_{i,j} \left| \boldsymbol{X}_j, A_j \right. \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Equation (3.13) is the standard textbook expression for a tobit model with heteroskedasticity (e.g. Wooldridge (2010)).

Of course, since the number of analysts is endogenous, it is reasonable to expect that $\xi_{i,j}$ in Equation (3.7) will not be mean-independent of A_j. Thus, we will again use the mergers of brokerage houses as an instrument for analyst coverage. Any remaining unobservable correlation between the latent demand factor $\xi_{i,j}$ and the latent consideration factor $\epsilon_{i,j}$ will be assumed away. That is in line with manager i conducting the stock evaluation independently of random factors that form his consideration set.

3.2.3. Fund Returns from Analyst Coverage

Based on Equations (3.10) and (3.13), manager i's expected total return (net from transaction costs) from investing optimally in a stock j that belongs to his consideration set, $\mathcal{J}_i$, is:

$$\mathbb{E}\left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}\right) w_{i,j} - \frac{1}{2} w_{i,j}^2 \left| d_{i,j} = 1, \boldsymbol{X}_j, A \right.\right] = \frac{1}{2} \mathbb{E}\left\{\left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}\right)^+\right]^2 \left| \boldsymbol{X}_j, A \right.\right\} \tag{3.14}$$

Consequently, *before* manager i's consideration set is formed, the unconditional expected total return from any available stock $j \in \mathcal{J}$ is:

$$\mathbb{E}\left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}\right) w_{i,j} - \frac{1}{2} w_{i,j}^2 \,|\boldsymbol{X}_j, A\right] = \frac{1}{2}\pi_{i,j}\mathbb{E}\left\{\left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \xi_{i,j}\right)^+\right]^2 |\boldsymbol{X}_j, A\right\} \tag{3.15}$$

Based on Equation (3.15), the following proposition is proved in the Appendix 3.9:

Proposition 1. *Manager i's unconditional expected total return from optimally investing in stock j is monotonically increasing in stock's analyst coverage, A_j.*

According to Proposition 1, any holdings that a fund manager has in stocks with high coverage should outperform any holdings in stocks with low coverage. This proposition provides the basis for the return analysis that we conduct in Section 6.

3.3. Data

3.3.1. Mutual Fund Stock Holdings

Our stock holdings data refer to the active equity funds in the US and are drawn from the CDA/Spectrum Mutual Fund Common Stock Holdings database provided by Thomson Reuters. The database sources from semi-annual mandatory filings and quarterly voluntary disclosures to the SEC by mutual funds. We merge the CDA/Spectrum database with the survivorship-bias free CRSP mutual fund database. The CRSP mutual fund database provides information on a variety of mutual fund characteristics such as assets under management and fund investment objectives. We exclude sector funds, since their consideration set is predetermined by the industry in which they specialize.

Specifically, in order to keep only actively managed, non-sector domestic equity funds in our sample, we apply the following detailed screening procedures. First, to exclude international, bond and index funds, we require (1) funds' investment

objective code reported by CDA/Spectrum to be aggressive growth, growth or growth and income, (2) their investment objectives in CRSP to be equity (E) and domestic (D) at the first two levels, (3) their CRSP objectives not to be EDCL, which indicates S&P500 index fund, and (4) their names not to contain the word "index". Second, to exclude sector funds, we require funds' CRSP investment objectives at the third level to be either (C) or (Y). Third, to exclude the possible presence of hedge funds, we require funds' CRSP investment objectives not to be (H) or (S) at the last level.

Overall, our sample is an unbalanced panel from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3), with observations at a quarterly frequency. Although our censored model is parametric, we identify the distribution of the manager-specific portfolio parameters non-parametrically, by performing the estimation at the fund level. To obtain stable estimates in our mixture structure, we include in our analysis only funds that hold at least 20 stocks.[14]

Applying these filters leads us to a sample of 1,708 unique funds for the whole period. On average, in every quarter, there are 585 mutual funds and 4,208 stocks with a complete list of characteristics (described below). We construct a fund's portfolio weight on a stock by dividing the dollar value of the stock holding by the fund's assets under management. The latter are defined to be equal to the total dollar value of the fund's stock holdings.[15] We also calculate the mutual fund ownership of a stock (i.e. the fraction of a firm's equity in the mutual fund portfolios), dividing the total number of a company's shares held by mutual funds by the respective total number of shares outstanding.

We present the summary statistics of the mutual fund stock holdings in Panel A of Table 3.1. The average portfolio value of a fund is 1.2 billion dollars, while the median is 194 million dollars. The standard deviation of the portfolio value is 5.8

[14]We also did experiments with alternative thresholds (e.g. 30 and 40 stocks). Since large funds invest in many stocks, the assets under management-weighted averages of the estimates were virtually the same.

[15]Active equity funds rarely invest in other asset classes.

billion dollars. The average portfolio weight of a fund on a stock is 3.2 basis points (bps), while the standard deviation is 30.3 bps. Of course, in line with the observed portfolio sparsity, the median portfolio weight is zero. On average, a fund holds 89 stocks during a quarter. The median number of stocks that are held is 51, while the standard deviation is about 157. As far as the mutual fund ownership of a stock is concerned (i.e. MFO, defined as the ratio of the sum of shares held by mutual fund managers over the total number of shares outstanding), it is on average 5%. The corresponding standard deviation is 7%, while the median mutual fund ownership is 1%.

3.3.2. Stock Characteristics

Our data on the analyst coverage of stocks are obtained from the Institutional Brokers Estimates System (IBES) Database at a monthly frequency. Monthly data on stock prices and returns are drawn from CRSP, while the accounting variables are collected from Compustat at a quarterly frequency. Stocks with missing observations are dropped from our sample.

We only focus on ordinary common shares traded at NYSE, AMEX, and Nasdaq. Analyst coverage (i.e. AC) is the number of analysts who make an earnings per share prediction for the forthcoming (financial) year in a given month. We keep stocks with zero coverage. The financial characteristics that we examine are market capitalization (i.e. Size), book to market ratio (i.e. Book to Market), turnover ratio (i.e. Turnover, defined as volume over number of outstanding shares), momentum (i.e. Momentum, defined as the past annual return) and volatility (i.e. Volatility, defined as the standard deviation of monthly returns in the past year). In order to merge the stock characteristics data with the Mutual Funds Holdings Database of Thompson Reuters, we average their values in every quarter. We assume that a manager's investment decision in quarter t is based on the stocks' analyst coverage

in that quarter and the above risk factors in quarter $t-1$.

The summary statistics of stock characteristics are depicted in Panel B of Table 3.1. The average number of analysts covering a stock is 3.82, while the median number is 1. The standard deviation of analyst coverage is 6.22. The average price of a stock during our sample period is 20 dollars, with a standard deviation of 41. Turning to the financial characteristics, the average size of firm is 2 billion dollars, while the median is 143 million dollars. Book to Market is on average 0.74 with a standard deviation 2.49. The average monthly turnover ratio is 11% with a corresponding standard deviation of 0.19. Momentum is on average 19% with a standard deviation of 2. Volatility is on average 14% and has a standard deviation of 11%.

3.4. Maximum Likelihood Estimation

3.4.1. Log Likelihood

Thus far, our assumptions regarding the distributions of the latent factors, namely that (i) $\epsilon_{i,j}\left|\boldsymbol{X}_j, A_j \overset{i.i.d}{\sim} Logistic\left(0,1\right)\right.$, so that $\pi_{i,j}=\exp\left(\kappa_i+\lambda_i A_j\right)/\left(1+\exp\left(\kappa_i+\lambda_i A_j\right)\right)$, (ii) $\xi_{i,j}\left|\boldsymbol{X}_j, A_j \sim \mathcal{N}\left(0,\sigma_{i,j}^2\right)\right.$ with $\sigma_{i,j}=\exp(\gamma_i+\delta_i A_j)$, and (iii) that $\epsilon_{i,j}$ and $\xi_{i,j}$ are independent, imply that manager i's log-likelihood is:

$$\begin{aligned}\mathcal{L}_i &= \sum_0 \log\left[\left(1-\pi_{i,j}\right)+\pi_{i,j}\left[1-\Phi\left(\frac{\alpha_i+\boldsymbol{\beta_i}\boldsymbol{X}_j}{\sigma_{i,j}}\right)\right]\right] \\ &+\sum_1 \log\left\{\pi_{i,j}\sigma_{i,j}\phi\left[\frac{w_{i,j}-\left(\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j\right)}{\sigma_{i,j}}\right]\right\}\end{aligned} \tag{3.16}$$

where $\Phi\left(\cdot\right)$ and $\phi\left(\cdot\right)$ are respectively the cumulative and probability density function of the standard normal distribution, $\sum_0$ is the sum over the stocks that manager i does not hold (i.e. the set $\{j\in\mathcal{J}_i:\ w_{i,j}=0\}$) and $\sum_1$ is sum over the stocks in which manager i invests (i.e. the set $\{j\in\mathcal{J}_i:\ w_{i,j}>0\}$).

The intuition behind the derivation of the likelihood is quite simple. The reason for which manager i does not invest in stock j is either because he does not pay attention to the stock or because he pays attention to the stock, but evaluates its growth prospects to be not that profitable (according to his fund's factor model). Both mutually exclusive cases imply that manager i's latent factors are not high enough, so that the probability of such an event is $(1-\pi_{i,j})+\pi_{i,j}(1-\Phi_{i,j})$. On the other hand, manager i's positive holdings on stock j necessitate attention and their actual level is determined by the factor model and the associated trading costs. Consequently, the density (kind of "probability") in that case is given by the product of the probability that manager i's latent consideration factor was high enough for the stock to be in his consideration set, $\pi_{i,j}$, and the conditional density of the positive portfolio weight *given* consideration, $\sigma_{i,j}\phi_{i,j}$.

All in all, our framework belongs to the class of censored mixture models, and in particular constitutes a logistically-mixed tobit model with heteroskedasticity. In the extreme case scenario in which manager i considers in his portfolio choice every existing stock (i.e the whole set $\mathcal{J}$), $\pi_{i,j}=1$ and the model reduces to a standard heteroskedastic Tobit model.

3.4.2. Coefficient Estimates

We estimate the model for every mutual fund in every period of our sample. Consistent with the notion that larger funds matter more for the industry, we focus in every period only on the *assets under management-weighted averages* of the fund parameters. For comparison, we also estimate two additional restricted models that our framework nests (henceforth referred to as AC Full Model). The first one is a version in which analyst coverage impacts only the extensive margin (henceforth referred to as AC Extensive Model), while the second one is a version in which analyst coverage impacts only the intensive margin (henceforth referred to as AC Intensive Model). We

present their time series averages in Panels A of Tables 3.2, 3.3 and 3.4 (along with standard errors of the mean in parenthesis). To assess the stability of the parameters over time, we also divide our sample into two roughly equal subsamples, using the fourth quarter of 1998 (1998Q4) as a break point.

As reported in Column 1 of Panel A in Table 3.2, the coefficient of analyst coverage at the extensive margin in the AC Full Model is on average 1.55. Based on the standard error of the mean, its 95% confidence interval is $[1.47, 1.62]$. According to Columns 2 and 3 of the same table, the coefficient takes higher values in the first subsample than in the second one, i.e. the 95% confidence interval changes from $[1.50, 1.74]$ to $[1.39, 1.56]$ as we move from Subsample 1 to Subsample 2. Moreover, according to Panel A of Table 3.3, the coefficient is virtually the same when we shut down the impact of analyst coverage at the intensive margin. Its average is slightly reduced in Subsample 1 (where it becomes 1.58 from 1.62) and slightly increased in Subsample 2 (where it becomes 1.52 from 1.47), as reported in Columns 2 and 3 respectively. On the other hand, the coefficient of analyst coverage at the intensive margin in the AC Full Model has an overall mean of 0.02 with a very tight 95% confidence interval. Its value in Subsample 2 is slightly higher than in Subsample 1 (i.e. 0.0275 vs. 0.0183). According to Panel A of Table 3.4, when we shut down the effect of analyst coverage on the extensive margin, the coefficient at the intensive margin is moderately increased to 0.03 for the whole sample period (as stated in Column 1). Again, the 95% confidence interval is quite tight. In both subsamples, its value is similar (i.e. 0.033 and 0.036 respectively).

Turning to the coefficients of the portfolio risk factors, their estimates seem to be similar across the three different models. On average, mutual funds have persistent preferences for stocks with large market capitalization and low book to market ratio (i.e. overvalued stocks). Their values in the whole sample are about 0.01 and -0.01 respectively. In Subsample 2, the coefficient of Log Size is decreased approximately

by a factor of 8 compared to its value in Subsample 1. It also changes from highly statistically significant to statistically insignificant. On the other hand, the coefficient of Book to Market displays more stable properties. Its values in the Subsamples 1 and 2 are -0.012 and -0.014 respectively. Their t-statistics have an absolute value of more than 6 in both cases.

Furthermore, mutual funds tend also to prefer stocks with high turnover ratio in Subsample 1. The opposite is true in Subsample 2. In whole sample, the AC Full Model yields that the turnover ratio coefficient is 0.003 and not statistically significant at a 5% significance level. In the same spirit, mutual funds invest in stocks with high momentum in Subsample 1 and in stocks with low momentum in Subsample 2. According to the AC Full Model, the 95% confidence interval for the coefficient of momentum in the whole sample is $[-0.007, -0.002]$. As for the stocks' volatility, mutual funds have always an aversion to it. In all cases, its coefficient is about -0.04.

3.4.3. Model Fit

According to the estimation results of the AC Full Model, the coefficients of analyst coverage in the extensive and the intensive margin are statistically significant, so that analysts affect equity demand through both the direct marketing and the precision channel. To buttress this finding, we also perform nested likelihood ratio tests for hypotheses according to which analysts impact on only one margin. The results of these tests are presented in Table 3.5.

In the first row of the table, we test the null hypothesis $H_0 : \delta_i = 0$ (AC Extensive Model) against the alternative hypothesis $H_1 : \delta_i \neq 0$ (AC Full Model). Under the null, analyst coverage affects the extensive margin but not the intensive margin. The test is performed by calculating the test statistic $D = 2[\log(L_{H_1}) - \log(L_{H_0})]$ (i.e., twice the difference between the log-likelihood in the AC Full Model and the log-likelihood in the AC Extensive Model) for every manager in every period. We then

take assets under management-weighted averages of the D-statistic across managers to derive the test statistic for whole cross-section in that period. Subsequently, we average over time the test statistics of every period to obtain the value of the D-statistic in respective sample period. The calculated values are remarkably high for the whole sample (e.g. 190.33) as well as for the subsamples (e.g. 34.03 and 348.97 in Subsample 1 and 2 respectively). To test H_0, we compare these figures to critical values of the χ^2 distribution with 1 degree of freedom. We reject the null hypothesis at any reasonable value of statistical significance, confirming that analyst coverage has a statistical significant effect on the intensive margin of portfolio choice.

In the same spirit, we test the null hypothesis $H_0 : \lambda_i = 0$ (AC Intensive Model) against the alternative $H_1 : \lambda_i \neq 0$ (AC Full Model) in the second row of the table. According to this null, analyst coverage matters for the extensive margin but not for the extensive. The value of the D-statistic is now about 60 in the whole sample as well as in the two subsamples. This figure is again much higher than critical values of the χ^2 distribution with 1 degree of freedom, at any reasonable significance level. Hence, H_0 is rejected, showing that analyst coverage has a statistical significant effect on the extensive margin as well.

3.4.4. Analyst Coverage Economic Effect

Our logistically-mixed tobit model with heteroskedasticity implies that manager i's expected portfolio weight on any available stock, given its characteristics is:

$$\begin{aligned} \mathbb{E}\left(w_{i,j}\,|\boldsymbol{X}_j, A_j\right) &= \underbrace{\pi_{i,j}\Phi\left(\frac{\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j}{\sigma_{i,j}}\right)}_{\mathbb{P}(w_{i,j}|P_j,\boldsymbol{X}_j,A_j)}\underbrace{\left[(\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j)+\sigma_{i,j}\frac{\phi\left(\frac{\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j}{\sigma_{i,j}}\right)}{\Phi\left(\frac{\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j}{\sigma_{i,j}}\right)}\right]}_{\mathbb{E}(w_{i,j}|w_{i,j}>0,\boldsymbol{X}_j,A_j)} \qquad (3.17) \\ &= \pi_{i,j}\left[(\alpha_i+\boldsymbol{\beta}_i\boldsymbol{X}_j)\,\Phi_{i,j}+\sigma_{i,j}\phi_{i,j}\right] \end{aligned}$$

Consequently, the marginal effect with respect to analyst coverage is:

$$\frac{\partial \mathbb{E}\left(w_{i,j} \,|\boldsymbol{X}_j, A_j\right)}{\partial A_j} = \underbrace{\pi_{i,j}\left(1-\pi_{i,j}\right)\lambda_i\left[\left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j\right)\Phi_{i,j} + \sigma_{i,j}\phi_{i,j}\right]}_{\text{Extensive Margin}} + \underbrace{\pi_{i,j}\phi_{i,j}\sigma_{i,j}\delta_i}_{\text{Intensive Margin}} \tag{3.18}$$

Equation (3.18) comprises a decomposition of the analysts' contribution to a manager's expected portfolio weight on a stock into two distinctive components corresponding to the extensive and intensive margin. In particular, the first term is the marginal effect of analyst coverage under the null hypothesis $H_0 : \delta_i = 0$; this is the marginal effect of the AC Extensive Model, in which analysts coverage impact only the extensive margin. In contrast, the second term is the marginal effect of analyst coverage under the null hypothesis $H_0 : \lambda_i = 0$; that is the marginal effect of the AC Intensive Model, in which analysts coverage matters only for the intensive margin.

Since our model is nonlinear, we evaluate the marginal effect of analyst coverage on manager i's portfolio in a given period at the average contemporaneous stock characteristics. Subsequently, we take assets under management-weighted averages to derive the marginal effect of the period. The corresponding *economic effect* is then defined to be the marginal effect of the period times the contemporaneous standard deviation of analyst coverage. Averaging that figure over time yields the economic effect of analyst coverage in the whole sample.

Our estimation results regarding the economic effect of analyst coverage are presented in Panels B of Tables 3.2, 3.3 and 3.4. According to the AC Full Model in Table 3.2, the total economic effect of analyst coverage in the whole sample is 0.88 bps with a t-statistic of 7.04. About 0.67 bps (or 76%) are attributed to the extensive margin, while 0.21 bps (or 24%) are attributed to the intensive margin. Given that the average portfolio weight of a mutual fund on a stock is 3.2 bps, our estimates provide evidence according to which analysts affect equity demand mainly from the

direct marketing channel of their research. The extensive vs. intensive margin decomposition of the AC economic effect in the two subsamples is very similar. In terms of magnitude, in Subsample 1, the total economic effect in Subsample 1 is 0.94 bps with a t-statistic of 4.32. In Subsample 2, the total economic effect is 0.83 bps with a t-statistic of 6.64

For reasons of comparison, we also estimate the economic effect of the analysts in the AC Extensive Model and in the AC Intensive Model. As we report in Table 3.2, we find that in the AC Extensive Model the economic effect for the whole sample is 0.58 bps, which is lower than the economic effect at the extensive margin in the AC Full Model. In particular, it is 0.54 bps in Subsample 1 and 0.63 bps in Subsample 2. On the other hand, the economic in the AC Intensive Model for the whole sample is 0.25 bps, i.e. slightly higher than the economic effect at the intensive margin in the AC Full Model. Its value is virtually the same in both subsamples.

3.5. Analyst Coverage Instrument

3.5.1. Mergers of Brokerage Firms

To identify latent factors correlated with analyst coverage, we invoke the Difference-in-Difference (DID) regression model of Hong and Kacperczyk (2010) for mergers of brokerage firms. For every merger m in the sample period, we specify a time window of one year before and one year after the quarter in which the event takes place. We index with $\tau = 0$ the pre-event period and with $\tau = 1$ the post-event period, so that the pair (m, τ) corresponds to a quarter in our data. We let $Merge_{m,\tau}$ be a dummy variable equal to one if merger m has happened. We also define the dummy variable $Affected_{j,m}$, which indicates whether stock j is affected by merger m. This is the case if one of the brokerage firms that is merged covers stock j in the pre-event period.

Consequently, the DID regression equation is:

$$\begin{aligned} A_{j,m,\tau} \quad = \quad & b_0 + b_1 Merge_{m,\tau} + b_2 Affected_{j,m} + \\ & b_3 Merge_{m,\tau} \times Affected_{j,m} + \boldsymbol{b_4 Controls}_{j,m,0} + \omega^A_{j,m,\tau} \end{aligned} \tag{3.19}$$

where the vector $\boldsymbol{Controls}_{j,m,0}$ includes financial characteristics of stock j in the pre-event period of merger m as well as merger and industry fixed effects. Importantly, the error term, $\omega^A_{j,m,\tau}$, represents unobservable characteristics of stock j that are relevant for its coverage. We refer to this term as *residual coverage*. As in any standard DID framework, $\omega^A_{j,m,\tau}$ is assumed to satisfy the property of mean independence (given all the regressors in the equation) and have the same distribution over time.

The point from incorporating Equation (3.19) in our model is that, naturally, some of the elements that residual coverage encompasses overlap with the latent consideration and demand factors of the managers. To this end, Equation (3.19) is useful not only because it identifies the treatment effect of mergers on analysts (b_3), but also because it produces estimates of the residual coverage in the post-merge period (the difference between actual and predicted coverage when $\tau = 1$). With these estimates in hand, we are able to proxy for the correlations between A_j and the latent factors in our demand system.

Indeed, both the latent consideration and demand factors, $\epsilon_{i,j}$ and $\xi_{i,j}$, can be decomposed into a manager-specific projection on the residual coverage, $\omega^A_{j,m,1}$, and an independent by construction residual latent term (i.e., $\tilde{\epsilon}_{i,j}$ or $\tilde{\xi}_{i,j}$ respectively):

$$\epsilon_{i,j} = \mathbb{E}_i \left(\epsilon_{i,j} \left| \omega^A_{j,m,1} \right. \right) + \tilde{\epsilon}_{i,j} \tag{3.20}$$

$$\xi_{i,j} = \mathbb{E}_i \left(\xi_{i,j} \left| \omega^A_{j,m,1} \right. \right) + \tilde{\xi}_{i,j} \tag{3.21}$$

The two projections on residual coverage can then be approximated through poly-

nomials and eventually become additional factors in our demand system, under the premise that they control for the analysts' endogeneity.[16] That is, Equations (3.1) and (3.11) are respectively replaced by:

$$d_{i,j} = \mathbf{1}\left[\kappa_i + \lambda_i A_j + \psi_i^{ext}\left(\omega_j^A\right) + \tilde{\epsilon}_{i,j} \geq 0\right] \tag{3.22}$$

$$w_{i,j} = \left(\alpha_i + \boldsymbol{\beta}_i \boldsymbol{X}_j + \psi_i^{int}\left(\omega_j^A\right) + \tilde{\xi}_{i,j}\right)^+ \tag{3.23}$$

where $\tilde{\epsilon}_{i,j}$ and $\tilde{\xi}_{i,j}$ are assumed to have the same distributional restrictions as the original latent factors. In that way, we recast the issue of possible correlations between the latent factors and analyst coverage as an omitted-variable bias problem of the original model.

3.5.2. DID Estimation

Using the brokerage house mergers' data of Hong and Kacperczyk (2010), we identify 11 quarters during the period 1984-2005 in which a merger took place.[17] The results of our DID analysis are presented in Table 3.6. t-statistics are calculated based on standard errors clustered at the stock level. Moreover, stocks are grouped into industries using the 17-industry classification of Fama and French, available from Kenneth R. French's website.[18]

Column 1 contains the estimated coefficients of Equation (3.19). In particular, we see that the average treatment effect of a merger on the analyst coverage of a stock is a reduction by 0.65. This change, which is highly statistically significant with a t-stat of 4.8, is consistent with the fact that around 1 analyst gets fired every time

[16] The polynomial functions are in principle non zero, unless $\epsilon_{i,j}$ and $\xi_{i,j}$ are independent of $\omega_{j,m,1}^A$.

[17] In fact, there are two instances in which more than one merger occurred in the same quarter (i.e. two mergers took place in 1998Q1 and three in 2000Q4). In those cases, we treated the coinciding mergers as a unique event.

[18] 1http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

two brokerage houses merge.

To get a better picture of how analysts impact equity demand in the aggregate, we also conduct a DID merger analysis for a stock's mutual fund ownership (MFO) and price. Specifically, in Column 2, we see that the treatment effect of a merger on MFO is a statistical significant reduction by about 1.34%. This effect is quite sizable given that the MFO is on average 5% (recall the summary statistics in Table 3.1). As for the impact on the price, in Column 3, we see that a merger results in a statistical significant price decrease of approximately 3.35 dollars, which translates into a drop by 17% given that the average price is 19.67 dollars. Hence, the magnitude of these reduced form results suggest that analyst coverage is an important determinant of the aggregate equity demand in the mutual fund industry (i.e, firing 1 analyst reduces a stock's MFO by approximately 2.06% and price by 26.2%).

3.5.3. Control Function Estimation

The results from estimating the AC Full Model by accounting for the endogeneity of analyst coverage are presented in Table 3.7. The averages over time correspond now to 11 quarters during the period 1985-2005. A brokerage houses merger took place one year before each of these quarters. As a benchmark, in Column 1 of Panel A, we list the average coefficient estimates from the new sample period in the uncorrected AC Full Model of Section 4. The associated economic effects are depicted in Panel B. The estimated coefficient of AC in the extensive margin is 2.12 and implies an economic effect of 0.82 bps. On the other hand, the estimated coefficient of AC in the intensive margin is 0.02, entailing an economic effect of 0.21 bps. Overall, the total economic effect of analysts is estimated to be 1.03 bps; 80% of the effect is extensive, while 20% is intensive.

The control function estimation results based on the mergers' DID residual coverage are presented in Columns 2 and 3 of the same table. In particular, in Column

2, the approximation of the control functions in the two margins, $\psi_i^{ext}(\cdot)$ and $\psi_i^{int}(\cdot)$, is linear. The estimated coefficient of AC in the extensive margin is decreased to 0.81 (i.e. roughly 62%). The associated economic effect on the extensive margin also decreases to 0.40 bps (i.e. it is 51% lower). In contrast, the coefficient of AC in the intensive margin is increased to 0.026 (i.e. approximately 30%), as is the implied economic effect, which is now 0.275 bps (i.e. about 31% higher). The total economic effect of AC is estimated to be lower than without the control functions. Specifically, it becomes 0.68 bps (i.e. about 34% lower). The extensive effect still dominates the intensive effect but to a lesser extent than before, accounting for approximately 60% of the total AC impact.

The results are qualitatively similar, when we quadratically approximate the control functions in Column 3 of the above table. The coefficient of AC in the extensive margin is estimated to be equal to 0.73, which higher than in the linear case, but still lower than without correcting for the AC endogeneity. The coefficient of AC in the intensive margin is estimated to be equal to 0.27, which is slightly higher than in the linear control functions' case. The economic effect of the extensive margin is now 0.56 bps. The economic effect on the intensive margin is now decreased to 0.18 bps. That is, there is a -14% correction in the intensive margin relative to the benchmark case which does not control for AC's endogeneity. In total, the economic effect decreases to 0.74 bps, and roughly 76% of it is extensive.

Consequently, correcting for the endogeneity of analyst coverage with control functions of mergers' residual coverage amounts to roughly decreasing the economic effect by 31% (to about 0.71 bps). The direct marketing channel through of analyst coverage continues to dominate the intensive margin either by a 60-40% or a 76-24% proportion.

3.6. Robustness

3.6.1. Analyst EPS Forecasts and Stock Recommendations

Thus far, we have shown the impact of the analysts on mutual fund portfolios through the channels of directing marketing at the extensive margin and signal precision enhancement at the intensive margin. Yet, in principle, analysts might also be affecting managers' expectations through their stock earnings forecasts and buy/sell recommendations (e.g., **?**, **?**, **?**). To allow for this possibility, we re-estimate the AC Full Model treating the analyst earnings per share (EPS) forecasts and stock recommendations as additional financial characteristics of stock j. In particular, in the vector of stock j's factors, $\boldsymbol{X}_j$, we add in either stock j's consensus EPS forecast or stock j's consensus recommendation. In either case, to account for analysts' different opinions (e.g., Diether, Malloy, and Scherbina (2002), Chan, Hwang, and Mian (2005)), we also include the standard deviation of the analysts' EPS forecasts or recommendations.

Data on the mean and standard deviation of the analysts' EPS forecasts (for the forthcoming fiscal year) and buy/sell recommendations are extracted from IBES at a monthly frequency and averaged for every quarter.[19] For the purpose of the exercise, stocks with zero analyst coverage are dropped from the estimation. We also set the standard deviation of stocks covered by only one analyst to zero. The summary statistics of the consensus and standard deviation of the analysts' EPS forecasts and buy/sell recommendations in a quarter are presented in Appendix Table 3.1. The average consensus EPS forecast is 1.57 dollars with a standard deviation of 2.86. The average standard deviation of the analysts' EPS forecasts is 0.13, while the median is 0.06. The average consensus recommendation is 2.23 with a standard deviation of 0.61. Analyst recommendations are measured on a scale of 1 to 5, with 1 denoting a strong suggestion to buy and 5 a strong suggestion to sell. The average standard

[19] The data on the analysts' recommendations are available from IBES only after the third quarter of 1993.

deviation of the buy/sell recommendations is 0.63, while the median is 0.71.

Analyst's EPS Forecasts

Appendix Table 3.2 presents the estimation results of the AC Full Model with Analyst's EPS Forecasts. Panel A shows that, for the whole sample period, the average coefficient of analyst coverage at the extensive margin is 0.5178 with a t-statistic of 14.44. In Subsample 1, it equals 0.2965 (with a t-statistic of 10.52), while in Subsample 2, it equals 0.7424 (with a t-statistic of 13.76). On the other hand, the coefficient of analyst coverage for the whole sample period is estimated to be 0.0184 with a t-statistic of 24.46. It equals 0.0153 in Subsample 1 and 0.0215 in Subsample 2 (with t-statistics of 14.4 and 23.42 respectively).

The economic effects of analyst coverage are depicted in Panel B of the same table. In the whole sample, the full effect of analyst coverage is 1.1024 bps, with a t-statistic of 10.37. About 62.41% (i.e. 0.688 bps with a t-statistic of 7.19) is attributed to the extensive margin, while 37.59% (i.e. 0.4144 bps with a t-statistic of 14.32) is due to the intensive margin. In Subsample 1, the full effect of analyst coverage is 1.3512 bps, with 56.85% vs. 43.15% extensive vs. intensive margin decomposition. As for the AC economic effect in Subsample 2, it is estimated to be 0.849 bps, with a proportion of 71.44% (i.e. 0.6065 bps) being attributed to the extensive margin and 28.57% (i.e. 0.2426 bps) being attributed to the intensive margin.

Comparing the above results with the estimation results of Table 3.2, we see that the inclusion of the analysts' EPS forecasts decreases the coefficients of analyst in both margins, but increases the full AC effect. Moreover, the economic effect of analyst coverage is also strengthened, as is its proportion against the extensive margin. As expected, we also see that in Panel of Table 3.2 the consensus of the analysts' EPS forecasts enters in the intensive margin positively (i.e. 0.006 for the whole sample), while the standard deviation enters negatively (i.e. -0.0148 for the whole sample).

Both are statistically significant in the whole sample period as well as in the two subsamples. The coefficients of the other financial controls have similar signs with those in Table 3.2.

To account for the potential endogeneity of analyst coverage, in Appendix Table 3.3 we further present the estimation results with control functions of residual coverage. Once again, we focus only on the 11 quarters one year after the mergers of brokerage houses. Without correcting for the AC endogeneity, the coefficient of the analysts at the extensive margin is estimated to be 0.3553, with a corresponding economic effect of 0.4001 bps. When we include linear control functions of residual coverage in the estimation, the coefficient of AC at the extensive margin increases to 0.4652 and the corresponding economic effect rises to 0.5336 bps. With quadratic control functions, the coefficient of AC at the extensive margin further rises to 0.72, but the corresponding economic effect further decreases to 0.6632 bps.

Turning to the AC coefficient in the intensive margin, in the uncorrected case, it is equal 0.0174 with a corresponding economic effect of 0.3045 bps. With linear control functions, the coefficient of the analysts increases to 0.025 and the economic effect to 0.3283 bps. With quadratic control function, the coefficient of the analysts in the intensive margin becomes 0.023 and the economic effect 0.2851 bps. As for the extensive vs. intensive margin decomposition, it is 56.78% vs. 43.22% in the uncorrected case. When the control functions are linear, it changes to 61.91% vs. 38.09%. With quadratic control functions it becomes 70.26% vs. 29.73%. Moreover, in the uncorrected case, the full economic effect of the analysts is 0.7046 bps. In the linear control function case, the full effect increases to 0.8619 bps, while it becomes 0.9438 bps in the quadratic control function case.

Analyst Stock Recommendations

In Appendix Table 3.4, we present the estimation results of the AC Full Model with analysts' Buy/Sell recommendations as additional financial characteristics in the intensive margin. Since the IBES data on the analyst stock recommendations start in the fourth quarter of 1993, we now split our sample into two equal subsamples using the third quarter of 2004 (2004Q3). Column 1 in Panel A shows that the AC coefficient at the extensive margin is 0.5403 (with a t-statistic of 18.57), while the AC coefficient at the intensive margin is 0.0264 (with a t-statistic of 17.12). The total economic effect of the analysts is estimated to be equal to 0.6116 bps, with 58.15%-41.85% extensive vs. intensive margin proportion.

In Subsample 1, the coefficient of the analysts at the extensive margin is 0.6061 with a corresponding economic effect of 0.3021 bps. On the other hand, the coefficient of the analysts in the intensive margin is 0.0228 with a corresponding economic effect of 0.2612 bps. In Subsample 2, the extensive margin AC coefficient is 0.473 and its implied economic effect is 0.4092 bps. As for the intensive margin AC coefficient in the second subsample, it equals 0.03, entailing an economic effect of 0.2508 bps.

A comparison of the aforementioned results with the ones of Table 3.2 yields that the inclusion of the analyst stock recommendations moderates the total economic effect of the analysts, particularly at the extensive margin. However, despite its decrease, the economic effect of the analysts at the extensive margin continues to be sizable and dominant. As expected, the coefficient of the consensus recommendation is negative (i.e. recall that high ranking denotes a suggestion to sell), as is the coefficient of the standard deviation of the analyst recommendations that captures the analysts' disagreement.

In Appendix Table 3.5, we present the estimation results in the AC Full Model with analysts' Buy/Sell recommendations and control functions of residual coverage

(following the mergers of brokerage houses[20]) to correct for the AC endogeneity. Without any correction, the coefficient of the analysts at the extensive margin is 0.6264, with an implied economic effect of 0.404 bps. With linear control functions of residual coverage, the AC extensive margin coefficient decreases to 0.5763, with an implied economic effect of 0.419 bps (i.e. roughly the same as before). With quadratic control functions, the coefficient of the number of analysts at the extensive margin further decreases to 0.4897 and has a smaller economic effect of 0.3618 bps.

On the other hand, in the uncorrected case, the coefficient of analyst coverage at the intensive margin is estimated to be equal 0.0193, with an implied economic effect of 0.3399 bps. With linear control functions, the AC intensive margin coefficient increases to 0.0239, but the economic effect decreases to 0.3014 bps. With quadratic control functions, the coefficient of the number of analysts at the intensive margin becomes equal to 0.021, with a corresponding economic effect of 0.3326 bps (as in the uncorrected case).

Therefore, accounting for the endogeneity of analyst coverage in the AC Full Model with the analyst stock recommendations peters down the full AC economic effect, from 0.7439 bps in the uncorrected case to 0.7204 in the linear control function case, and 0.6945 bps in the quadratic control function case. This decrease is attributed to a reduced AC effect at the extensive margin, which again is still prevalent. That is, the change in the AC extensive vs. intensive margin decomposition is from 54.30%-45.70% in the uncorrected case to 58.16%-41.83% in the linear control function case, and 52.10%-47.90% in the quadratic control function case.

[20]Instead of 11, we now refer to 9 quarters that followed the merger of a brokerage house. Again, this is because the data on the analyts' recommendations in IBES start in the fourth quarter of 1993.

3.6.2. Financial Characteristics in the Extensive Margin

We also relax the assumption according to which stock j's financial characteristics, $\boldsymbol{X_j}$, affect manager i's portfolio choice only through the intensive margin. Indeed, thus far, we have assumed that manager i uses the financial risk factors only to evaluate the expected returns of the stocks that he considers. However, it could be possible that some financial characteristics contribute to the formation of manager i's consideration set, $\mathcal{J}_i$, in the first place, just as the analysts. We investigate the consequences of this possibility for the estimation of the AC economic effect in Appendix Tables 3.6 and 3.7. In particular, without taking a stance on which financial characteristics play that dual role, we allow to all of them to potentially affect the extensive margin. For brevity, we refer to this augmented specification as the AC-FC Full Model (where the FC stands for financial characteristics).

Appendix Table 3.6 presents the estimation results from the mixed tobit estimation of AC-FC Full Model using the whole sample period and without correcting for the endogeneity of AC. Specifically, according to Column 1 of Panel A, the estimated coefficient of number of analysts at the extensive margin is 1.42 with a t-statistic of 26.2. Columns 2 and 3 show that its value in Subsamples 1 and 2 is about the same (i.e 1.44 and 1.40 respectively). On the other hand, the coefficient of AC at the intensive margin is 0.025 for the whole sample period, while 0.021 and 0.029 in Subsample 1 and 2, respectively. As in the AC Full Model (of Table 3.2), the 95% confidence interval for that coefficient is quite tight. More importantly, Panel B shows that the economic effect of AC and its extensive vs. intensive margin decomposition is not that different from the AC Full Model. Specifically, for the whole sample, the economic impact is 1.13 bps with a t-statistic of 8. About 81% of the effect (i.e. 0.93 bps) comes from the extensive margin, while 19% (i.e. 0.20 bps) comes from the intensive margin. In Subsample 1, the total economic effect is 0.96 bps and about 73% of it can be attributed to the extensive margin. That proportion increases to

88% in Subsample 2, where the total economic effect is 1.31.

In Appendix Table 3.7, we display the results from employing the control function method to correct for the endogeneity of AC using the mergers of brokerage houses. Column 1 of Panel A shows that the coefficient of AC in the 11 post-merge quarters of the sample is 1.36 at the extensive margin and 0.026 at the intensive margin. According to Panel B, the associated economic effect is 1.18 bps with 0.64% (i.e. 0.75 bps) coming from the extensive margin and 0.36% (i.e. 0.43 bps) coming from the intensive margin. In Column 2, where the control functions in both margins are linear functions of residual coverage, the coefficient of AC in the extensive margin drops to 0.77 bps, while the coefficient in the intensive margin remains approximately the same as before (i.e. 0.027). As a result, the extensive margin economic effect decreases by 8% (i.e. to 0.69 bps), while the intensive margin drops only by 6% compared to the uncorrected case (i.e. to 0.40 bps). The total economic effect thus drops to 1.09 bps (i.e. a 8% reduction). The extensive vs. intensive margin decomposition becomes 62% vs. 48%. In Column 3, we approximate the control functions also quadratically. The estimated coefficient of AC in the extensive margin is 1.02, i.e. higher than in the case in which the control function is linear, yet still lower than in the uncorrected case. On the other hand, the coefficient of coverage in the intensive margin is equal to 0.30 and thus higher than before. The total economic effect now becomes 0.69 bps which is smaller than the linear control function case. About 60% of the effect (i.e. 0.41 bps) is attributed to the extensive margin, while 40% (i.e. 28 bps) is attributed to the intensive margin.

In a nutshell, even if some financial characteristics impact the extensive margin, the estimated economic effect of analyst coverage and its decomposition are not that different from the ones in our baseline specification.

3.7. Mutual Fund Returns from Analyst Coverage

In this section, we conduct a portfolio strategy to test the predictive power of analyst coverage for returns of mutual fund stock holdings based on our Proposition 1, i.e. fund managers' holdings in stocks with high analyst coverage should outperform their holdings in stocks with low coverage.

Based on the distribution of a stock's number of analysts in a given quarter, we make a high or low coverage stock distinction. Specifically, any stock whose number of analysts is higher than the third quartile (Q3) of the AC distribution is characterized as a high coverage stock. On the other hand, any stock whose number of analysts is lower than the first quartile (Q1) of the AC distribution is considered to be a stock with low coverage. When we calculate the quartiles of the AC distribution, we do not take into account stocks with zero coverage. However, we do keep stocks without coverage in all the performance tests that we conduct. As a result, the AC cutoff for a stock to be characterized as a low coverage one ranges from 1 to 3, while the AC cutoff for a stock to be characterized as a high coverage on ranges from 8 to 13.[21]

Having established the two sets of stocks to be compared, we collect their monthly returns in the prospective quarter. Consistent with the fact that the mutual fund holdings data are available at a quarterly frequency, high or low AC stocks in quarter t are matched with their monthly returns in quarter $t+1$. We focus on both on the raw returns of the high coverage stocks (vs. the low ones) as well as their Carhart 4-factor adjusted alpha.[22] To assess the stability of our return tests over time, we again divide our 1982Q1-2015Q3 sample period into two subsamples using 1998Q4 as a break point. We also compute t-statistics using Newey-West standard errors with a lag order of 3 months.

[21]The performance results are very similar for alternative neighboring cutoffs (e.g. the 20% vs. the 80% percentile or the 30% vs. 70% percentile).

[22]The Carhart four factors are obtained from Kenneth French's website at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

In Table 3.8, we present the performance results from unconditionally value-weighting the monthly returns of these stocks. According to Column 1 of Panel A, the average monthly return from a strategy that longs the high AC portfolio and at the same time sorts the low AC portfolio is .168% with a t-stat of 2.37. As Column 1 of Panel B shows, this return has an alpha of 21% with a t-statistic of 2.71. In the subsamples, the mean monthly raw returns are similar, i.e. .144% in Subsample 1 and .192% in Subsample 2. The t-statistic in Subsample 2 is lower than in Subsample 1, i.e. 1.58 vs. 1.98. The alphas are .146% in Subsample 1 and .286% in Subsample 2. The corresponding t-stat in now higher in Subsample 2 than in Subsample 1, i.e. 2.83 vs. 1.74.

Next, we follow the methodology of Coval and Moskowitz (2001) and Cohen, Frazzini, and Malloy (2008) to test whether mutual funds earn higher returns from stocks with high coverage. In particular, we look at every mutual fund in quarter t and find which high or low coverage stocks it holds and at which proportion. Although not that often, it is possible that some funds will be investing exclusively in high or low coverage stocks. Once we identify a mutual fund's weights in high or low coverage stocks, we rescale them so that they add up to 1 for both the high AC and the low AC portfolio. Subsequently, we obtain the monthly returns of the two portfolios in the next quarter. Once we have all the mutual funds' returns in the high and low AC portfolios, we weight them by their assets under management to derive the return of the high and low AC portfolio from the perspective of mutual funds.

The results from this exercise are present in Table 3.9. As stated in Column 1 of Panel A, the average return from a strategy that longs the high AC portfolio and sorts the low AC portfolio in the same way that mutual funds do yields is .328% per month with a t-stat of 2.07. The alpha from that strategy is .39% with a t-statistic of 2.74. In Subsample 1, the average monthly raw return is .449% with a t-statistic of 2.2. In Subsample 2, the average monthly raw return is lower, e.g. .20%, as is the

t-stat 0.87. However, in both subsamples the alphas are similar in terms of estimates as well as t-statistics. Specifically, the alpha in Subsample 1 is .351% with a t-statistic of 2.21, while in Subsample 2 it is .398% with a t-statistic if 1.78. The results from the two above performance tests suggest that the raw returns and the alpha of the high coverage stocks are magnified (approximately doubled) under the mutual fund holdings. We can then do a back-of-the-envelope calculation, whereby we can take the 60-40 estimate obtained from our extensive-intensive margin decomposition and apply it to the 5% per annum alpha. We retrieve a 3% per annum alpha or benefit from coverage on the extensive margin and 2% per annum alpha for the intensive margin.

3.8. Estimates of the Marginal Cost of Producing Information

3.8.1. Banks' Stock Coverage Decisions

Having estimated the stock portfolio demand for each manager in the mutual fund industry, we can now recover the marginal cost of analyst coverage that banks face when they decide how many analysts to allocate for the coverage of a given stock.

Letting $\mathcal{I}$ be the set of all managers in the mutual fund industry, the market value of stock j that the mutual fund industry demands at the aggregate level equals the sum over all managers of the product of the portfolio weight that each manager places on that stock and the value of his total assets, i.e:

$$D_j\left(A_j, \boldsymbol{X}_j, \{\xi_{i,j}\}_{i\in\mathcal{I}}\right) = \sum_{i\in\mathcal{I}} w_{i,j}\left(A_j, \boldsymbol{X}_j, \xi_{i,j}\right) W_i \tag{3.24}$$

Taking the expectation over the managers' latent demand shocks, $\{\xi_{i,j}\}_{i\in\mathcal{I}}$, conditional on stock j's analyst coverage, A_j, and financial characteristics, $\boldsymbol{X_j}$, yields that

the expected aggregate demand for stock j is:

$$\begin{aligned} D_j^*\left(A_j, \boldsymbol{X}_j\right) &\equiv \mathbb{E}\left[D_j\left(A_j, \boldsymbol{X}_j, \{\xi_{i,j}\}_{i\in\mathcal{I}}\right)|A_j, \boldsymbol{X}_j\right] \\ &= \sum_{i\in\mathcal{I}} \mathbb{E}\left(w_{i,j}\,|A_j, \boldsymbol{X}_j\right) W_i \end{aligned} \tag{3.25}$$

where manager i's expected portfolio weight on stock j is given by Equation (3.17).

The objective of the investment banks is to maximize the profits from executing the transactions in which the mutual fund industry is expected to engage for each stock. We assume that the revenue that the sell-side makes for stock j is a constant fraction, k_j, of stock j's expected aggregate demand, $D_j^*(A_j, \boldsymbol{X}_j)$. Additionally, we assume that the cost for covering stock j is convex and specifically quadratic, with a marginal cost coefficient m_j. Therefore, investment banks collectively solve the following expected profit maximization problem:

$$\max_{\{A_j\geq 0\}_{j\in\mathcal{J}}} \left\{\sum_{j\in\mathcal{J}}\left[k_j D_j^*\left(A_j, \boldsymbol{X}_j\right) - \frac{1}{2} m_j \left(1+A_j\right)^2\right]\right\} \tag{3.26}$$

The KKT conditions entail that:

$$\begin{aligned} \frac{\frac{\partial D_j^*(A_j, \boldsymbol{X}_j)}{\partial A_j}}{1+A_j} &= \frac{m_j}{k_j} \quad \text{if } A_j > 0 \\ \frac{\partial D_j^*\left(A_j, bmX_j\right)}{\partial A_j} &< \frac{m_j}{k_j} \quad \text{if } A_j = 0 \end{aligned} \tag{3.27}$$

From Equation (3.27), we obtain either a value or a lower bound for the *relative cost-benefit ratio* of information of stock j, which we define by the ratio $\frac{m_j}{k_j}$. The critical value for determining that ratio is the increase in stock j's expected demand due to the increase in its coverage by one extra analyst, $\frac{\partial D_j^*(A_j, bmX_j)}{\partial A_j}$, divided by the total number of analysts who cover stock j plus one.

3.8.2. Correlations with Stock Characteristics

The correlations of a stock's relative cost-benefit ratio of information with its financial characteristic are depicted in Table 3.10. The correlation with LogSize is about 0.07 with a t-statistic of 8.1 and the correlation with BTM is approximately -0.14 with a t-statistic of -4.9. Therefore, for large firms or value firms (with low book-to-market ratio) the relative cost-benefit ratio from producing information is higher.

The correlation with Turnover is about 0.06 with a t-statistic of 3.4, while the correlation with Momentum is about 0.12 with a t-statistic of 5.5. Hence, stocks with higher turnover and momentum have a higher relative cost-benefit ratio of information. On the other hand, the correlation with Volatility is about 0.01 with a t-statistic of 1.9, so that high volatility stocks have only a small, positive and marginally statistically insignificant correlation with the relative cost-benefit ratio of information.

The magnitude and statistical significance of these correlations are in general robust when we divide our sample to two equal subsamples. One exception is Volatility, since in Subsample 1 its correlation with the relative cost-benefit ratio of information has a t-statistic of 1.3, while in Subsample 2 it has a t-statistic of 2.1. Moreover, although the statistical significance of Momentum is similar in both subsamples, its magnitude is about three times higher in Subsample 1 than in Subsample 2 (i.e., being equal to 0.18 instead of 0.06).

3.9. Conclusion

A significant part of the buy-side's equity research is outsourced to the sell-side, i.e. security analysts working for investment banks. Despite the lack of data on fund-level expenses for this research, we can nonetheless quantify its benefits by modeling the impact of analyst coverage on fund portfolio weights. We identify two distinct channels of sell-side research, direct marketing versus signal precision, by comparing

the impact of coverage on the extensive versus intensive margins of fund portfolio weights. Using stock holdings data of US active equity funds, we find that analyst coverage significantly increases a manager's portfolio weight on a stock. Moreover, we find that this increase is due to both the extensive and intensive margins. We then complete our estimate of the benefits of sell-side research by showing that the return performance of mutual fund holdings are positively impacted by analyst coverage.

Table 3.1: Summary Statistics

This table reports the summary statistics of all the variables in our sample. Panel A refers to the stock holdings of the mutual funds. Portval is the portfolio value of a mutual fund. Portwt is the portfolio of a mutual fund on a stock. Numst is the number of different stocks that a mutual fund holds. MFO is the percentage of mutual fund ownership of a stock, computed as the ratio of the sum of shares held by mutual fund managers over the total number of shares outstanding. Panel B shows the stock characteristics. All variables are constructed by averaging the respective monthly values in a given quarter. AC (i.e. Analyst Coverage) is the average number of analysts who make an earnings per share prediction for the forthcoming (fiscal) year in a given quarter. Price is a stock's price. Size is market capitalization. Book to Market is the ratio of book value over market capitalization. Turnover is the ratio of the traded volume over number of outstanding shares. Momentum is the past annual return. Volatility is the standard deviation of monthly returns in the past year. The sample period is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3).

Panel A: Mutual Fund Stock Holdings

	Mean	S.D.	Median	Min	Max
Portval ($ mil.)	1,239.171	5,752.573	194.16	0.04	364,034.9
Portwt	3.2 bps	30.3 bps	0	0	0.97
Numst	88.56	156.93	51	20	3,079
MFO	0.05	0.07	0.01	0.00	1

Panel B: Stock Characteristics

	Mean	S.D.	Median	Min	Max
AC (Analyst Coverage)	3.82	6.22	1	0	55
Price	19.67	41.38	12.79	0.02	4613.33
Size ($ mil.)	2,085.571	11,852.72	143.21	0.09	729,152.8
Log Size	18.91	2.2	18.79	11.37	27.32
Book to Market	0.74	2.49	0.56	-667.54	188.33
Turnover	0.11	0.19	0.06	0.00	31.48
Momentum	0.19	1.98	0.01	-1.00	404.71
Volatility	0.14	0.11	0.12	0.00	4.09

Table 3.2: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model, in which the number of analysts impacts both the extensive and the intensive margin. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the first quarter of 1982 (1982Q1) until the fourth quarter of 1998 (1998Q4). The third column presents the estimates in Subsample 2, i.e. from the first quarter of 1999 (1999Q1) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
AC Extensive	1.5465	1.6201	1.4719
	[40.46]	[26.17]	[34.52]
Log Size	0.0050	0.0086	0.0013
	[3.84]	[9.96]	[0.56]
Book to Market	-0.0132	-0.012	-0.0143
	[-10.06]	[-10.27]	[-6.11]
Turnover	0.003	0.0165	-0.0107
	[1.04]	[3.52]	[-4.72]
Momentum	-0.0044	0.0019	-0.0108
	[-3.04]	[2.22]	[-4.18]
Volatility	-0.0400	-0.0364	-0.0437
	[-13.32]	[-6.79]	[-16.87]
AC Intensive	0.0229	0.0183	0.0275
	[24.86]	[22.27]	[18.89]

Panel B: AC Economic Effects (bps)			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Full Effect	0.8849	0.9435	0.8254
	[7.04]	[4.32]	[6.64]
Extensive Effect	0.6739	0.7127	0.6345
	[10.54]	[10.76]	[5.65]
Intensive Effect	0.2110	0.2308	0.1909
	[8.91]	[5.56]	[8.09]

Table 3.3: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Extensive Model

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Extensive Model, in which the number of analysts impacts only the extensive margin. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i and α_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the first quarter of 1982 (1982Q1) until the fourth quarter of 1998 (1998Q4). The third column presents the estimates in Subsample 2, i.e. from the first quarter of 1999 (1999Q1) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
AC Extensive	1.5456	1.5751	1.5156
	[37.65]	[22.46]	[36.22]
Log Size	0.0109	0.0118	0.0100
	[76.46]	[91.94]	[48.92]
Book to Market	-0.0091	-0.0097	-0.0085
	[-13.8]	[-8.72]	[-12.27]
Turnover	0.0096	0.0212	-0.0021
	[3.75]	[4.55]	[-3.73]
Momentum	-0.0004	0.002	-0.0028
	[-1.05]	[3.7]	[-10.89]
Volatility	-0.0355	-0.0354	-0.0357
	[-11.62]	[-6.33]	[-14.8]
AC Intensive	-	-	-
	[-]	[-]	[-]

Panel B: AC Economic Effects (bps)			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Full Effect	0.5777	0.6164	0.5384
	[7.6]	[4.72]	[7.07]
Extensive Effect	0.5777	0.6164	0.5384
	[7.6]	[4.72]	[7.07]
Intensive Effect	-	-	-
	[-]	[-]	[-]

Table 3.4: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Intensive Model

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Intensive Model, in which the number of analysts impacts only the intensive margin. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the first quarter of 1982 (1982Q1) until the fourth quarter of 1998 (1998Q4). The third column presents the estimates in Subsample 2, i.e. from the first quarter of 1999 (1999Q1) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
AC Extensive	-	-	-
	[-]	[-]	[-]
Log Size	0.0068	0.0089	0.0048
	[5.95]	[17.31]	[2.14]
Book to Market	-0.0091	-0.009	-0.0093
	[-8.02]	[-11.12]	[-4.33]
Turnover	0.0118	0.0273	-0.0039
	[4.41]	[6.76]	[-1.77]
Momentum	-0.0062	-0.0031	-0.0094
	[-4.87]	[-6.4]	[-3.8]
Volatility	-0.0494	-0.0512	-0.0477
	[-20.49]	[-12.9]	[-17.6]
AC Intensive	0.0345	0.0332	0.0357
	[50.08]	[42.26]	[32.09]

Panel B: AC Economic Effects (bps)			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Full Effect	0.2544	0.2543	0.2546
	[5.41]	[11.96]	[2.77]
Extensive Effect	-	-	-
	[-]	[-]	[-]
Intensive Effect	0.2544	0.2543	0.2546
	[5.41]	[11.96]	[2.77]

Table 3.5: Nested Likelihood Ratio Test for Restricted Versions of the Full Demand Model of Analyst Coverage

This table presents nested likelihood ratio tests for shutting down the analysts' impact on either the intensive or the extensive margin. The first row refers to testing the null hypothesis $H_0 : \delta_i = 0$ (AC Extensive Model) against the alternative hypothesis $H_1 : \delta_i \neq 0$ (AC Full Model). Under this null, coverage affects only the extensive margin. The second row refers to testing the null hypothesis $H_0 : \lambda_i = 0$ (AC Intensive Model) against the alternative hypothesis $H_1 : \lambda_i \neq 0$ (AC Full Model). Under this null, coverage affects only the intensive margin. In every period, the test statistic $D = 2[log(L_{H_1}) - log(L_{H_0})]$ is calculated (which is twice the difference between the log-likelihood in the AC Full Model and the log-likelihood in the AC Extensive/Intensive Model) is calculated for every manager. Then, the D-statistics are averaged across the managers in the period, using their assets under management as weights. Subsequently, the test statistics from every period are averaged over time to obtain the D-statistic in the sample. Column 1 reports the average D-statistic in the whole sample. Column 2 reports the average D-statistic in Subsample 1. Column 3 reports the average D-statistic in Subsample 2. The presented average D-statistics can be compared with the critical values of the χ^2 distribution with 1 degree of freedom, e.g. 10.83 at the $\alpha = 0.1\%$ level of statistical significance. The sample period is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The fourth quarter of 1998 (1998Q4) is used as the breaking point of the two subsamples.

Panel A: Average D-Statistic			
	(1)	(2)	(3)
Hypothesis Test	Whole Sample	Subsample 1	Subsample 2
Extensive vs. Full Model	190.33	34.03	348.97
Intensive vs. Full Model	59.52	59.75	59.28

Table 3.6: Difference in Difference Estimation From Merging Brokerage Houses

This table presents three difference in difference regression results from mergers of brokerage houses. The dependent variable in the fist column is analyst coverage (AC). The dependent variable in the second column is the mutual fund ownership of a stock (MFO). The dependent variable in the third column is the price of a stock (Price). Given the quarter in which a merger takes place, observations are obtained one year before (pre-event period) and one year after (post-event period). The indicator variable Merge denotes the pre-event period when it is equal to zero and the post-event period when it is equal to one. The indicator variable Affected equals one if a stock is covered by both merging brokerage houses and zero otherwise. Accounting characteristics from the pre-event period are included as controls. The "L" in front of them denotes the lag operator. AC is an additional control in the MFO and Price regressions. All three regressions have merger fixed effects and industry fixed effects based on the 17-industry classification of Fama and French. t-statistics based on standard errors clustered at the stock level are shown in brackets. The regressions refer to 11 quarters during the period 1983-2005 in which a merger took place. The list with the exact dates of the mergers is available in Hong and Kacperczyk (2010).

	Change in		
	(1) AC	(2) MFO	(3) Price
Merge	0.3292	0.0081	0.2346
	[15.81]	[14.99]	[1.75]
Affect	3.6386	-0.0089	1.9695
	[11.24]	[-2.82]	[2.51]
Merge x Affect	-0.6511	-0.0134	-3.3512
	[-4.83]	[-4.07]	[-4.39]
L(AC)		0.0041	-0.1138
		[24.18]	[-1.465]
L(Log Size)	2.2602	-0.0016	7.070
	[31.24]	[-2.75]	[20.52]
L(Book to Market)	0.0764	-0.0114	-0.7895
	[0.73]	[-5.56]	[-1.70]
L(Turnover)	4.5514	0.0807	5.8801
	[8.470]	[9.130]	[4.191]
L(Momentum)	-0.2054	0.0001	0.4458
	[-3.25]	[0.66]	[1.82]
L(Volatility)	-2.2161	-0.0194	-38.9436
	[-3.14]	[-1.71]	[-9.46]
Merger FE	YES	YES	YES
Industry FE	YES	YES	YES
Observations	63,416	63,416	63,416
R^2	0.41	0.20	0.26

Table 3.7: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model With Control Function for Demand Shocks Correlated with Analyst Coverage

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model, correcting for the endogeneity of AC through the control function method. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Column 1 presents the estimates in the AC Full Model, without correcting for the endogeneity of analyst coverage. Columns 2 and 3 present the estimates in the AC Full Model correcting for the endogeneity of AC with two control functions (CF's) of residual coverage; one for the extensive and one for the intensive margin. In Column 2, the control functions are approximated linearly. In Column 3, the control functions are approximated quadratically. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. The sample period corresponds to 11 quarters from the period 1985-2005; one year before each quarter a merger of brokerage houses took place.

Panel A: Coefficient Estimates			
	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
AC Extensive	2.1217	0.8097	0.7331
	[13.32]	[9.05]	[11.93]
Log Size	0.0039	0.0044	-0.0003
	[0.91]	[2.19]	[-0.05]
Book to Market	-0.0192	-0.0172	-0.0208
	[-3.22]	[-4.78]	[-3.72]
Turnover	-0.0135	-0.0121	-0.0145
	[-1.72]	[-1.71]	[-3.09]
Momentum	0.0029	-0.0014	-0.0066
	[0.56]	[-0.56]	[-1.26]
Volatility	-0.0218	-0.0142	-0.0119
	[-1.89]	[-1.24]	[-0.73]
AC Intensive	0.0203	0.0259	0.0272
	[7.1]	[16.87]	[12.83]

Panel B: AC Economic Effects (bps)			
	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
Full Effect	1.0303	0.6801	0.7367
	[3.24]	[10.3]	[4.15]
Extensive Effect	0.8203	0.4047	0.561
	[2.62]	[5.79]	[2.94]
Intensive Effect	0.2100	0.2753	0.1757
	[5.93]	[5.13]	[6.43]

Table 3.8: Unconditional Value-Weighted Returns of Stocks with High Coverage vs. Stocks with Low Coverage

This table presents the results from the unconditionally value-weighted returns test of high analyst coverage stocks vs. low analyst coverage stocks. Returns are collected on a monthly basis and measured in basis points (i.e. bps). t-statistics are computed using Newey-West standard errors with lag order 3 are shown in brackets. In a given quarter, high coverage stocks are stocks whose analyst coverage is higher than the 3rd quartile of the AC distribution. The four Fama-French factors are used. The sample period is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3).

Panel A: Raw Returns			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Mean	0.168	0.144	0.192
	[2.37]	[1.98]	[1.58]

Panel B: Returns Regression			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Mktrf	-0.027	0.0507	-0.110
	[-1.08]	[2.07]	[-3.22]
SMB	-0.171	-0.170	-0.152
	[-7.59]	[-4.84]	[-5.57]
HML	-0.144	-0.188	-0.0884
	[-4.14]	[-5.02]	[-2.29]
Momentum	0.0512	-0.0079	0.0484
	[1.93]	[-0.24]	[1.58]
Constant	0.210	0.146	0.286
	[2.71]	[1.74]	[2.83]

Table 3.9: Mutual Fund Value-Weighted Returns of Stocks with High Coverage vs. Stocks with Low Coverage

This table presents the results from the value-weighted returns test of high analyst coverage stocks vs. low analyst coverage stocks *conditional on mutual fund investment*. Returns are collected on a monthly basis and measured in basis points (i.e. bps). t-statistics are computed using Newey-West standard errors with lag order 3 are shown in brackets. In a given quarter, high coverage stocks are stocks whose analyst coverage is higher than the 3rd quartile of the AC distribution. The four Fama-French factors are used. The sample period is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3).

Panel A: Raw Returns			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Mean	0.328	0.449	0.206
	[2.07]	[2.19]	[0.87]

Panel B: Returns Regression			
	(1)	(2)	(3)
	Whole Sample	Subsample 1	Subsample 2
Mktrf	-0.0213	-0.0483	-0.00508
	[-0.71]	[-1.23]	[-0.10]
SMB	-0.610	-0.791	-0.525
	[-12.97]	[-12.98]	[-8.35]
HML	-0.146	-0.176	-0.142
	[-2.24]	[-2.46]	[-1.48]
Momentum	0.0693	0.0657	0.0631
	[1.79]	[1.25]	[1.22]
Constant	0.390	0.351	0.398
	[2.74]	[2.21]	[1.78]

Table 3.10: Correlations of the Relative Cost-Benefit of Information with Financial Characteristics

This table presents the correlations of the relative cost-benefit of information with financial stock characteristics [with t-statistics are shown in brackets]. The sample period is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3).

	Correlations of $\frac{\partial D_j^*(A_j, \boldsymbol{X_j})}{\partial A_j}/(1+A_j)$ with:				
	LogSize	BTM	Turnover	Momentum	Volatility
Whole Sample	0.072	-0.137	0.058	0.121	0.011
	[8.12]	[-4.91]	[4.78]	[5.48]	[1.92]
Subsample 1	0.068	-0.153	0.063	0.182	0.014
	[4.3]	[-3.28]	[3.41]	[4.57]	[1.32]
Subsample 2	0.077	-0.121	0.053	0.058	0.008
	[9.45]	[-3.97]	[3.38]	[3.89]	[2.14]

Appendix of Chapter 3

Proposition Proof

Under the optimal portfolio weight from Equation (3.13), the expected total excess return of investing in a stock j can be written as

$$\mathbb{E}\left[\left(\alpha_i + \boldsymbol{\beta}\boldsymbol{X}_j + \sigma_{i,j}\zeta_{i,j}\right)\left(\alpha_i + \boldsymbol{\beta}\boldsymbol{X}_j + \sigma_{i,j}\zeta_{i,j}\right)^+ - \frac{1}{2}\Big(\left(\alpha_i + \boldsymbol{\beta}\boldsymbol{X}_j + \sigma_{i,j}\zeta_{i,j}\right)^+\Big)^2 | \boldsymbol{X}_j, A_j\right] \tag{3.28}$$

where $\sigma_{i,j} = \exp\left(\gamma_i + \delta_i A_j\right)$ and $\zeta_{i,j} | \boldsymbol{X}_j, A_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. To simplify notations, let us denote $\alpha_i + \boldsymbol{\beta}\boldsymbol{X}_j = r_{i,j}$ and leave out the notation of conditioning on $\boldsymbol{X}_j, A_j$ as it is always there. Now this expectation term of (3.28) is equal to

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{2}\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j}\Big)^2 | \, r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\right)\mathbb{P}\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\Big) \\ &= \mathbb{E}\left(\frac{1}{2}\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j}\Big)^2 | \, r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\right)\Phi\Big(\frac{r_j}{\sigma_{i,j}}\Big)\pi_{i,j} \end{aligned} \tag{3.29}$$

To show that Equation (3.29) (i.e. the expected total excess return of investing in a stock j) is monotonically increasing in the analyst coverage A_j, we just need to show that its derivative with respect to A_j is strictly positive. The first conditional expectation term in (3.29) is

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{2}\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j}\Big)^2 | \, r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\right) \\ &= \frac{1}{2}\left[Var\Big(\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j}\Big) | \, r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\Big) + \mathbb{E}\Big(\Big(r_{i,j} + \sigma_{i,j}\zeta_{i,j}\Big) | \, r_{i,j} + \sigma_{i,j}\zeta_{i,j} > 0\Big)^2\right] \end{aligned} \tag{3.30}$$

where Var denotes variance. By the properties of truncated normal distributions

(truncated from below at 0), we have

$$Var\Big(\big(r_{i,j}+\sigma_{i,j}\zeta_{i,j}\big)\,|\,r_{i,j}+\sigma_{i,j}\zeta_{i,j}>0\Big)=\sigma_{i,j}^2\Bigg(1+\frac{\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)\phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}{1-\Phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}-\Big(\frac{\phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}{1-\Phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}\Big)^2\Bigg)$$
$$=\sigma_{i,j}^2\Bigg(1-\frac{\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}-\frac{\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}\Bigg)\quad(3.31)$$

and

$$\mathbb{E}\Big(\big(r_{i,j}+\sigma_{i,j}\zeta_{i,j}\big)\,|\,r_{i,j}+\sigma_{i,j}\zeta_{i,j}>0\Big)^2=\Bigg(r_{i,j}+\frac{\sigma_{i,j}\phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}{1-\Phi\big(-\frac{r_{i,j}}{\sigma_{i,j}}\big)}\Bigg)^2$$
$$=r_{i,j}^2+\frac{\sigma_{i,j}^2\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}+\frac{2r_{i,j}\sigma_{i,j}\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}\quad(3.32)$$

Hence

$$\mathbb{E}\Bigg(\frac{1}{2}\big(r_{i,j}+\sigma_{i,j}\zeta_{i,j}\big)^2\,|\,r_{i,j}+\sigma_{i,j}\zeta_{i,j}>0\Bigg)$$
$$=\frac{1}{2}\Bigg(\sigma_{i,j}^2\Bigg(1-\frac{\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}-\frac{\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}\Bigg)+\Bigg(r_{i,j}^2+\frac{\sigma_{i,j}^2\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}+\frac{2r_{i,j}\sigma_{i,j}\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}\Bigg)\Bigg)\quad(3.33)$$

Therefore, Equation (3.29) becomes

$$\mathbb{E}\Big(\frac{1}{2}\big(r_{i,j}+\sigma_{i,j}\zeta_{i,j}\big)^2\,|\,r_{i,j}+\sigma_{i,j}\zeta_{i,j}>0\Big)\Phi\big(\frac{r_j}{\sigma_{i,j}}\big)\pi_{i,j}$$
$$=\frac{1}{2}\Bigg(\sigma_{i,j}^2\Bigg(1-\frac{\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}-\frac{\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}\Bigg)+\Bigg(r_{i,j}^2+\frac{\sigma_{i,j}^2\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)^2}+\frac{2r_{i,j}\sigma_{i,j}\phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}{\Phi\big(\frac{r_{i,j}}{\sigma_{i,j}}\big)}\Bigg)\Bigg)\times\Phi\big(\frac{r_j}{\sigma_{i,j}}\big)\pi_{i,j}$$
$$=\frac{1}{2}\Bigg[\Phi\Big(\frac{r_{i,j}}{\sigma_{i,j}}\Big)(\sigma_{i,j}^2+r_{i,j}^2)+r_{i,j}\sigma_{i,j}\phi\Big(\frac{r_{i,j}}{\sigma_{i,j}}\Big)\Bigg]\pi_{i,j}\quad(3.34)$$

Now we want to show that the derivative of (3.34) above with respective to A_j is

strictly positive. Let

$$\frac{1}{2}\left[\Phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\left(\sigma_{i,j}^2+r_{i,j}^2\right)+r_{i,j}\sigma_{i,j}\phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\right]=V_{i,j} \tag{3.35}$$

such that Equation (3.34) can be written as $V_{i,j}\pi_{i,j}$. Then we want to show

$$\frac{\partial V_{i,j}\pi_{i,j}}{\partial A_j}=\pi_{i,j}\frac{\partial V_{i,j}}{\partial A_j}+V_{i,j}\frac{\partial \pi_{i,j}}{\partial A_j}>0 \tag{3.36}$$

Since we know $V_{i,j}>0$ always from its definition in Equation (3.34) (i.e. it is the expectation of a squared term multiplied by two probability terms so it is always positive), and

$$\frac{\partial \pi_{i,j}}{\partial A_j}=\lambda_i\pi_{i,j}(1-\pi_{i,j})>0 \tag{3.37}$$

based on the definition of $\pi_{i,j}$, we have the second term

$$V_{i,j}\frac{\partial \pi_{i,j}}{\partial A_j}>0 \tag{3.38}$$

Thus it is left to show that

$$\frac{\partial V_{i,j}}{\partial A_j}>0 \tag{3.39}$$

in the first term. Because

$$\frac{\partial V_{i,j}}{\partial A_j}=\frac{\partial V_{i,j}}{\partial \sigma_{i,j}}\frac{\partial \sigma_{i,j}}{\partial A_j}=\frac{\partial V_{i,j}}{\partial \sigma_{i,j}}\sigma_{i,j}\delta_i \tag{3.40}$$

we just need to derive the expression for $\frac{\partial V_{i,j}}{\partial \sigma_{i,j}}$. This derivative is

$$\begin{aligned}\frac{\partial V_{i,j}}{\partial \sigma_{i,j}}=&\frac{1}{2}\left[\left(\sigma_{i,j}^2+r_{i,j}^2\right)\phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\left(-\frac{r_{i,j}}{\sigma_{i,j}^2}\right)+2\Phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\sigma_{i,j}+r_{i,j}\left(\phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)+\phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\frac{r_{i,j}^2}{\sigma_{i,j}^2}\right)\right]\\ =&\frac{1}{2}\left[2\Phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\sigma_{i,j}\right]=\Phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\sigma_{i,j}\end{aligned} \tag{3.41}$$

Therefore,

$$\frac{\partial V_{i,j}}{\partial A_j} = \Phi\left(\frac{r_{i,j}}{\sigma_{i,j}}\right)\sigma_{i,j}^2\delta_i > 0 \tag{3.42}$$

Hence we have the desired result for (3.36). □

Tables of the AC Full Model with Analysts' EPS Forecasts and Buy/Sell Recommendations

Appendix Table 3.1: Summary Statistics of the Analysts' EPS Forecasts and Buy/Sell Recommendations

This table reports additional summary statistics for the Analysts' EPS Forecasts and Buy/Sell Recommendations. Consensus corresponds to the average and Standard Deviation to the standard deviation. The earnings per share (EPS) forecasts are expressed in dollars. The recommendations are measured on a scale of 1 to 5, with 1 denoting a strong suggestion to buy and 5 denoting a strong suggestion to sell. The sample period for the EPS forecasts is from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The sample period for the Buy/Sell Recommendations is from the fourth quarter of 1993 (1993Q4) until the third quarter of 2015 (2015Q3).

Analyst's EPS Forecasts & Buy/Sell Recommendations					
	Mean	S.D.	Median	Min	Max
Analyst EPS Forecasts					
Consensus	1.57	2.86	1.25	-909	152.67
Standard Deviation	0.13	1.32	0.06	0	717.17
Analyst Recommendations					
Consensus	2.23	0.61	2.2	1	5
Standard Deviation	0.63	0.4	0.71	0	2.83

Appendix Table 3.2: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model with Analysts' EPS Forecasts

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model with Analysts's EPS Forecasts, which incorporates the consensus (i.e. mean) and the standard deviation of the EPS forecasts made by analysts as additional financial characteristics in the *intensive* margin of the AC Full Model. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the first quarter of 1982 (1982Q1) until the fourth quarter of 1998 (1998Q4). The third column presents the estimates in Subsample 2, i.e. from the first quarter of 1999 (1999Q1) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
AC Extensive	0.5178	0.2965	0.7424
	[14.44]	[10.52]	[13.76]
Log Size	0.0091	0.0091	0.0092
	[13.92]	[7.19]	[29.58]
Book to Market	-0.0097	-0.011	-0.0083
	[-10.89]	[-7.15]	[-9.94]
Turnover	0.005	0.0128	-0.003
	[1.86]	[2.52]	[-4.4]
Momentum	-0.0018	-0.0001	-0.0036
	[-2.52]	[-0.05]	[-11]
Volatility	-0.0184	-0.0097	-0.0274
	[-6.08]	[-1.75]	[-14.65]
EPS Analyst Forecasts			
Consensus	0.0006	0.0008	0.0003
	[12.63]	[14.41]	[5.27]
Standard Deviation	-0.0148	-0.0143	-0.0152
	[-15.1]	[-9.69]	[-11.94]
AC Intensive	0.0184	0.0153	0.0215
	[24.46]	[14.4]	[23.42]

Panel B: AC Economic Effects (bps)

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
Full Effect	1.1024	1.3512	0.849
	[10.37]	[7.3]	[9.12]
Extensive Effect	0.688	0.7681	0.6065
	[7.19]	[4.58]	[6.79]
Intensive Effect	0.4144	0.583	0.2426
	[14.32]	[13.5]	[9.85]

Appendix Table 3.3: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model with Analysts' EPS Forecasts With Control Function for Demand Shocks Correlated with Analyst Coverage

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model with Analysts' EPS Forecasts, which incorporates the consensus (i.e mean) and the standard deviation of the EPS forecasts made by analysts as additional financial characteristics in the intensive margin of the AC Full Model, correcting for the endogeneity of AC through the control function method. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Column 1 presents the estimates in the AC Full Model with Analysts' EPS Forecasts, without correcting for the endogeneity of analyst coverage. Columns 2 and 3 present the estimates in the AC Full Model with Analysts' EPS Forecasts correcting for the endogeneity of AC with two control functions (CF's) of residual coverage; one for the extensive and one for the intensive margin. In Column 2, the control functions are approximated linearly. In Column 3, the control functions are approximated quadratically. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. The sample period corresponds to 11 quarters from the period 1985-2005; one year before each quarter a merger of brokerage houses took place.

Panel A: Coefficient Estimates

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
AC Extensive	0.3553	0.4652	0.7263
	[8.5]	[7.61]	[9.6]
Log Size	0.0112	0.0082	0.0087
	[34.2]	[10.78]	[19.34]
Book to Market	-0.0113	-0.0173	-0.0121
	[-4.15]	[-3.17]	[-4.08]
Turnover	-0.0046	-0.0108	-0.0046
	[-0.63]	[-1.11]	[-0.54]
Momentum	-0.0014	-0.0031	0.0016
	[-0.84]	[-0.73]	[1.25]
Volatility	-0.0066	-0.0019	0.0066
	[-0.7]	[-0.16]	[0.55]
EPS Analyst Forecast			
Consensus	0.0006	0.0005	0.0004
	[6.49]	[2.91]	[2.24]
Standard Deviation	-0.0201	-0.0259	-0.0207
	[-7.15]	[-5.05]	[-8.63]
AC Intensive	0.0174	0.025	0.023
	[8.54]	[12.8]	[12.49]

Panel B: AC Economic Effects (bps)

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
Full Effect	0.7046	0.8619	0.9483
	[5.35]	[6.8]	[5.29]
Extensive Effect	0.4001	0.5336	0.6632
	[2.49]	[4.9]	[2.59]
Intensive Effect	0.3045	0.3283	0.2851
	[2.96]	[8.95]	[6.71]

Appendix Table 3.4: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model with Analysts' Buy/Sell Recommendations

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model with Analysts's Buy/Sell Recommendations, which incorporates the consensus (i.e. mean) and the standard deviation of the Buy/Sell recommendations made by analysts as additional financial characteristics in the *intensive* margin of the AC Full Model. (The recommendations are measured on a 1-5 scale, with 1 denoting a strong recommendation to buy and 5 denoting a strong recommendation to sell.) The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the fourth quarter of 1993 (1993Q4) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the fourth quarter of 1993 (1993Q4) until the third quarter of 2004 (2004Q3). The third column presents the estimates in Subsample 2, i.e. from the fourth quarter of 2004 (2004Q4) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
AC Extensive	0.5403	0.6061	0.473
	[18.57]	[14.86]	[12.15]
Log Size	0.0068	0.0081	0.0053
	[7.8]	[7.95]	[3.82]
Book to Market	-0.0219	-0.0242	-0.0195
	[-6.13]	[-4.12]	[-4.9]
Turnover	-0.0146	-0.0196	-0.0095
	[-4.09]	[-3.31]	[-2.51]
Momentum	-0.004	0.0023	-0.0104
	[-1.1]	[0.39]	[-2.45]
Volatility	-0.0622	-0.058	-0.0664
	[-15.23]	[-9.52]	[-12.45]
Analyst Recom.			
Consensus	-0.0193	-0.0209	-0.0177
	[-5.44]	[-3.67]	[-4.23]
Standard Deviation	-0.0097	-0.0118	-0.0076
	[-2.69]	[-2]	[-1.85]
AC Intensive	0.0264	0.0228	0.03
	[17.12]	[10.69]	[14.42]

Panel B: AC Economic Effects (bps)

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
Full Effect	0.6116	0.5633	0.66
	[8.67]	[4.79]	[9.72]
Extensive Effect	0.3557	0.3021	0.4092
	[8.58]	[4.6]	[10.33]
Intensive Effect	0.2559	0.2612	0.2508
	[5.12]	[3.42]	[4.6]

Appendix Table 3.5: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC Full Model with Analysts' Buy/Sell Recommendations With Control Function for Demand Shocks Correlated with Analyst Coverage

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC Full Model with Analysts' Buy/Sell Recommendations, which incorporates the consensus (i.e mean) and the standard deviation of the Buy/Sell recommendations made by analysts as additional financial characteristics in the intensive margin of the AC Full Model, correcting for the endogeneity of AC through the control function method. (The recommendations are measured on a 1-5 scale, with 1 denoting a strong recommendation to buy and 5 denoting a strong recommendation to sell.) The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Column 1 presents the estimates in the AC Full Model with Analysts' Buy/Sell Recommendations, without correcting for the endogeneity of analyst coverage. Columns 2 and 3 present the estimates in the AC Full Model with Analysts' Buy/Sell Recommendations correcting for the endogeneity of AC with two control functions (CF's) of residual coverage; one for the extensive and one for the intensive margin. In Column 2, the control functions are approximated linearly. In Column 3, the control functions are approximated quadratically. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. The sample period corresponds to 9 quarters from the period 1995-2005; one year before each quarter a merger of brokerage houses took place.

Panel A: Coefficient Estimates

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
AC Extensive	0.6264	0.5763	0.4897
	[6.01]	[6.31]	[5.29]
Log Size	0.0062	0.0088	0.0075
	[3.15]	[7.09]	[6.51]
Book to Market	-0.017	-0.0315	-0.0127
	[-3.52]	[-2.53]	[-4.49]
Turnover	-0.0184	-0.0327	-0.0153
	[-3.97]	[-2.7]	[-4.92]
Momentum	-0.0011	-0.0012	0.0003
	[-0.23]	[-0.09]	[0.13]
Volatility	-0.0333	-0.0473	-0.0302
	[-4.2]	[-3.67]	[-4.35]
Analyst Recom.			
Consensus	-0.0157	-0.0305	-0.0116
	[-3.49]	[-2.42]	[-4.97]
Standard Deviation	-0.0081	-0.0067	-0.0019
	[-2.66]	[-2.29]	[-2.55]
AC Intensive	0.0193	0.0239	0.021
	[5.01]	[5.02]	[5.01]

Panel B: AC Economic Effects (bps)

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
Full Effect	0.7439	0.7204	0.6945
	[6.81]	[2.72]	[3.88]
Extensive Effect	0.404	0.419	0.3618
	[2.95]	[3.76]	[2.79]
Intensive Effect	0.3399	0.3014	0.3326
	[4.75]	[2.81]	[2.6]

Tables of the AC-FC Full Model

Appendix Table 3.6: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC-FC Full Model

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC-FC Full Model, in which the number of analysts *as well as financial characteristics* impact *both* the extensive and the intensive margin. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. In both panels, the first column presents the estimates from the whole sample period, i.e. from the first quarter of 1982 (1982Q1) until the third quarter of 2015 (2015Q3). The second column presents the estimates in Subsample 1, i.e. from the first quarter of 1982 (1982Q1) until the fourth quarter of 1998 (1998Q4). The third column presents the estimates in Subsample 2, i.e. from the first quarter of 1999 (1999Q1) until the third quarter of 2015 (2015Q3).

Panel A: Coefficient Estimates

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
AC Extensive	1.4222	1.4398	1.4044
	[26.20]	[18.25]	[18.86]
Log Size Ext.	-0.1513	-0.06	-0.244
	[-7.03]	[-1.97]	[-9.43]
Book to Market Ext.	0.185	-0.0012	0.374
	[1.46]	[-0.01]	[2.07]
Turnover Ext.	-2.077	-3.6228	-0.508
	[-7.56]	[-9.54]	[-1.74]
Momentum Ext.	-1.2301	-1.5377	-0.9178
	[-9.53]	[-7.43]	[-6.38]
Volatility Ext.	-4.4493	-4.924	-3.9676
	[-15.23]	[-10.84]	[-11.12]
Log Size Int.	0.0029	0.0055	0.0003
	[2.79]	[5.40]	[0.17]
Book to Market Int.	-0.0127	-0.0122	-0.0133
	[-8.54]	[-7.95]	[-5.17]
Turnover Int.	0.0117	0.0286	-0.0054
	[4.37]	[7.54]	[-2.24]
Momentum Int.	-0.0007	0.0053	-0.0067
	[-0.46]	[5.52]	[-2.65]
Volatility Int.	-0.0115	-0.0032	-0.02
	[-4.16]	[-0.68]	[-7.74]
AC Intensive	0.0251	0.0214	0.0289
	[31.33]	[24.88]	[24.19]

Panel B: AC Economic Effects (bps)

	(1) Whole Sample	(2) Subsample 1	(3) Subsample 2
Full Effect	1.1313	0.9595	1.3057
	[8.09]	[5.13]	[6.34]
Extensive Effect	0.9295	0.7044	1.158
	[6.75]	[3.89]	[5.67]
Intensive Effect	0.2018	0.2551	0.1478
	[18.00]	[13.67]	[18.32]

Appendix Table 3.7: Mixed Tobit Estimation of Mutual Fund Portfolio Choice in the AC-FC Full Model With Control Function for Demand Shocks Correlated with Analyst Coverage

This table presents the estimation results from three logistically-mixed Tobit models with heteroskedasticity of mutual fund portfolio choice in the AC-FC Full Model, correcting for the endogeneity of AC through the control function method. In the AC-FC Full Model, the number of analysts *as well as financial characteristics* impact *both* the extensive and the intensive margin. The results are time-series averages of the assets under management-weighted average estimates in every quarter. t-statistics based on the standard error of the mean are shown in brackets. Column 1 presents the estimates in the AC-FC Full Model, without correcting for the endogeneity of analyst coverage. Columns 2 and 3 present the estimates in the AC-FC Full Model correcting for the endogeneity of AC with two control functions (CF's) of residual coverage; one for the extensive and one for the intensive margin. In Column 2, the control functions are approximated linearly. In Column 3, the control functions are approximated quadratically. Panel A refers to the estimated parameters of the model. For clarity, the estimated values of the constants in the model (i.e. parameters κ_i, α_i and γ_i) are omitted. Panel B refers to the economic effect on the portfolio weight (in bps) of a 1 standard deviation increase in AC, calculated at the average stock characteristics of every period. It also decomposes the total economic effect into an extensive and an intensive margin component. The sample period corresponds to 11 quarters from the period 1985-2005; one year before each quarter a merger of brokerage houses took place.

Panel A: Coefficient Estimates

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
AC Extensive	1.357	0.7717	1.0249
	[10.79]	[8.00]	[6.30]
Log Size Ext.	-0.1756	0.0041	0.2111
	[-2.62]	[0.06]	[2.80]
Book to Market Ext.	0.0204	-0.6862	-0.4164
	[0.04]	[-1.82]	[-1.21]
Turnover Ext.	-2.9956	-1.5686	-2.0141
	[-4.52]	[-3.17]	[-2.63]
Momentum Ext.	-1.9321	-1.5774	-1.1076
	[-2.69]	[-3.04]	[-2.44]
Volatility Ext.	-2.1256	-3.5897	-3.6277
	[-2.81]	[-4.31]	[-3.41]
Log Size Int.	-0.0032	0.0063	0.0011
	[-0.49]	[3.3]	[0.22]
Book to Market Int.	-0.0191	-0.0078	-0.011
	[-2.3]	[-2.38]	[-2.53]
Turnover Int.	-0.0021	0.0088	-0.0007
	[-0.29]	[0.98]	[-0.13]
Momentum Int.	0.0016	0.0016	-0.0067
	[0.21]	[0.63]	[-1.42]
Volatility Int.	-0.0145	-0.0037	-0.0101
	[-1.35]	[-0.33]	[-0.49]
AC Intensive	0.0264	0.027	0.0303
	[9.46]	[19.43]	[10.80]

Panel B: AC Economic Effects (bps)

	Uncorrected	With Res. AC Control Function	
	(1)	(2)	(3)
CF's Approx.	-	Linear	Quadratic
Full Effect	1.1837	1.0918	0.6964
	[2.94]	[3.97]	[4.83]
Extensive Effect	0.7535	0.6884	0.4127
	[1.90]	[2.35]	[3.03]
Intensive Effect	0.4302	0.4034	0.2837
	[3.34]	[4.68]	[4.57]

Bibliography

Agarwal, S., J. C. Driscoll, X. Gabaix, and D. Laibson, 2009, "The Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation," *Brookings Papers on Economic Activity*, 2009(2), 51–117.

Aizawa, N., and Y. S. Kim, 2015, "Advertising and Risk Selection in Health Insurance Markets," *Working Paper*.

Arbel, A., S. Carvell, and P. Strebel, 1983, "Giraffes, Institutions and Neglected Firms," *Financial Analysts Journal*, 39(3), 57–63.

Bailey, W., A. Kumar, and D. Ng, 2011, "Behavioral Biases of Mutual Fund Investors," *Journal of Financial Economics*, 102(1), 1–27.

Bajari, P., and M. E. Kahn, 2005, "Estimating Housing Demand With an Application to Explaining Racial Segregation in Cities," *Journal of Business and Economic Statistics*, 23(1), 20–33.

Barber, B. M., and T. Odean, 2000, "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance*, 55(2), 773–806.

——— , 2002, "Online Investors: Do the Slow Die First?," *Review of Financial Studies*, 15(2), 455–487.

——— , 2008, "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors," *Review of Financial Studies*, 21(2), 785–818.

——— , 2013, "Chapter 22 - The Behavior of Individual Investors," vol. 2 of *Handbook of the Economics of Finance*, pp. 1533 – 1570.

Barber, B. M., T. Odean, and L. Zheng, 2005, "Out of Sight, Out of Mind: The Effects of Expenses on Mutual Fund Flows," *Journal of Business*, 78(6), 2095–2119.

Bayer, P., F. Ferreira, and R. McMillan, 2007, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 115(4), 588–638.

Bayer, P. J., B. D. Bernheim, and J. K. Scholz, 2009, "The effects of financial education in the workplace: Evidence from a survey of employers," *Economic Inquiry*, 47(4), 605–624.

Bishop, K. C., 2007, "A Dynamic Model of Location Choice and Hedonic Valuation," *Working Paper*.

Brandt, W. M., P. Santa-Clara, and R. Valkanov, 2009, "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns," *Review of Financial Studies*, 22(9), 3411–3447.

Branikas, I., H. Hong, and J. Xu, 2017, "Location Choice, Portfolio Choice," *NBER Working Paper*.

Campbell, J. Y., 2006, "Household finance," *The Journal of Finance*, 61(4), 1553–1604.

Carrillat, F. A., and A. d'Astous, 2012, "The Sponsorship-Advertising Interface: Is Less Better for Sponsors?," *European Journal of Marketing*, 46(3/4), 562–574.

Carter, D. M., 1996, *Keeping Score: An Inside Look at Sports Marketing*. Oasis Press.

Chan, K., C.-Y. Hwang, and G. M. Mian, 2005, "Mutual Fund Herding and Dispersion of Analysts' Earnings Forecasts," *Working Paper*.

Charles, K. K., E. Hurst, and N. Roussanov, 2009, "Conspicuous Consumption and Race," *The Quarterly Journal of Economics*, 124(2), 425–467.

Chemmanur, T. J., and A. Yan, 2009, "Advertising, Attention, and Stock Returns," *Working Paper*.

Chen, J., H. Hong, and J. C. Stein, 2002, "Breadth of ownership and stock returns," *Journal of Financial Economics*, 66(2), 171–205.

Cho, C.-H., and H. J. Cheon, 2004, "Why Do People Avoid Advertising on the Internet?," *Journal of Advertising*, 33(4), 89–97.

Cohen, L., A. Frazzini, and C. Malloy, 2008, "The Small World of Investing: Board Connections and Mutual Fund Returns," *Journal of Political Economy*, 116(5), 951–979.

Coval, J. D., and T. J. Moskowitz, 1999, "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance*, 54(6), 2045–2073.

Coval, J. D., and T. J. Moskowitz, 2001, "The Geography of Investment: Informed Trading and Asset Prices," *Journal of Political Economy*, 109(41), 811–841.

Cronqvist, H., 2006, "Advertising and Portfolio Choice," *Working Paper*.

Dahl, G. B., 2002, "Imobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70(6), 2367–2420.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, "Measuring mutual fund performance with characteristic-based benchmarks," *The Journal of finance*, 52(3), 1035–1058.

Danielson, M. N., 1997, *Home Team: Professional Sports and the American Metropolis*. Princeton University Press.

DeMarzo, P. M., R. Kaniel, and I. Kremer, 2004, "Diversification as a Public Good: Community Effects in Portfolio Choice," *Journal of Finance*, 59(4), 1677–1716.

Diamond, R., 2016, "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000," *American Economic Review*, 106(3), 479–524.

Diether, K. B., C. J. Malloy, and A. Scherbina, 2002, "Differences of Opinion and the Cross Section of Stock Returns," *Journal of Finance*, 57(5), 2113–2141.

Duflo, E., 2001, "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4), 795–813.

Euchner, C. C., 1993, *Playing the Field: Why Sports Teams Move and Cities Fight to Keep Them*. Johns Hopkins University Press.

Falkenstein, E. G., 1996, "Preferences for Stock Characteristics As Revealed by Mutual Fund Portfolio Holdings," *Journal of Finance*, 51(1), 111–135.

Fama, E. F., and K. R. French, 1992, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2), 427–465.

Fama, E. F., and K. R. French, 2010, "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *Journal of Finance*, 65(5), 1915–1947.

——— , 2015, "A Five-Factor Asset Pricing Model," *Journal of Financial Economics*, 116(1), 1 – 22.

Fang, L. H., J. Peress, and L. Zheng, 2014, "Does Media Coverage of Stocks Affect Mutual Funds'Trading and Performance?," *Review of Financial Studies*, 27(12), 3441–3466.

Fehle, F., S. Tsyplakov, and V. Zdorovtsov, 2005, "Can Companies Influence Investor Behaviour through Advertising? Super Bowl Commercials and Stock Returns," *European Financial Management*, 11(5), 625–647.

Feng, L., and M. S. Seasholes, 2008, "Individual Investors and Gender Similarities in An Emerging Stock Market," *Pacific-Basin Finance Journal*, 16, 44–60.

Focke, F., S. Ruenzi, and M. Ungeheuer, 2016, "Advertising, Attention, and Financial Markets," *Working Paper*.

French, K. R., 2008, "Presidential Address: The Cost of Active Investing," *Journal of Finance*, 63(4), 1537–1573.

French, K. R., and J. M. Poterba, 1991, "Investor Diversification and International Equity Markets," *American Economic Review*, 81(2), 222–226.

Gabaix, X., 2016, "Behavioral macroeconomics via sparse dynamic programming," working paper, National Bureau of Economic Research.

Gennaioli, N., and A. Shleifer, 2010, "What Comes to Mind," *Quarterly Journal of Economics*, 125(4), 1399–1433.

Goeree, M. S., 2008, "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 76(5), 1017–1074.

Goetzmann, W. N., and A. Kumar, 2008, "Equity Portfolio Diversification," *Review of Finance*, 12(3), 433–463.

Gomez, J.-P., R. Priestley, and F. Zapatero, 2009, "Implications of Keeping-Up-with-the-Joneses Behavior for the Equilibrium Cross Section of Stock Returns: International Evidence," *Journal of Finance*, 64(6), 2703–2737.

Gompers, P. A., and A. Metrick, 2001, "Institutional Investors and Equity Prices," *Quarterly Journal of Economics*, 116(1), 229–259.

Greenwood, R., and D. Scharfstein, 2013, "The Growth of Finance," *Journal of Economic Perspectives*, 27(2), 3–28.

Grinblatt, M., and M. Keloharju, 2001, "How Distance, Language, and Culture Influence Stockholdings and Trades," *Journal of Finance*, 56(3), 1053–1073.

Grinblatt, M., and S. Titman, 1989, "Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings," *Journal of Business*, 62(3), 393–416.

Grinblatt, M., S. Titman, and R. Wermers, 1995, "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior," *American Economic Review*, 85(5), 1088–1105.

Gârleanu, N., and L. H. Pedersen, 2013, "Dynamic Trading with Predictable Returns and Transaction Costs," *Journal of Finance*, 68(6), 2309–2340.

Grullon, G., G. Kanatas, and J. P. Weston, 2004, "Advertising, Breadth of Ownership, and Liquidity," *Review of Financial Studies*, 17(2), 439–461.

Heath, C., and A. Tversky, 1991, "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty," *Journal of Risk and Uncertainty*, 4(1), 5–28.

Heckman, J. J., 1977, "Sample Selection Bias as a Specification Error (with an Application to the Estimation of Labor Supply Functions)," *NBER Working Papers*.

Heckman, J. J., 1978, "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(4), 931–959.

Hillert, A., A. Kunzmann, and S. Ruenzi, 2017, "M & A(dvertising)," *Working Paper*.

Hizmo, A., 2015, "Risk in Housing Markets: An Equilibrium Approach," *Working Paper*.

Hjalmarsson, E., and P. Manchev, 2012, "Characteristic-Based Mean-Variance Portfolio Choice," *Journal of Banking and Finance*, 36(5), 1392–1401.

Hong, H., W. Jiang, N. Wang, and B. Zhao, 2014, "Trading for Status," *Review of Financial Studies*, 27, 3171–3212.

Hong, H., and M. Kacperczyk, 2010, "Competition and Bias," *Quarterly Journal of Economics*, 125(4), 1683–1725.

Hong, H., J. D. Kubik, and J. C. Stein, 2004, "Social Interaction and Stock-Market Participation," *Journal of Finance*, 59(1), 137–163.

Hong, H., T. Lim, and J. C. Stein, 2000, "Bad News Travels Slowly: Size, Analyst Coverage, and The Profitability of Momentum Strategies," *The Journal of Finance*, 55(1), 265–295.

Hong, H., and J. Xu, 2015, "Inferring Latent Social Networks from Stock Holdings," *Working Paper*.

Hortaçsu, A., and J. Kastl, 2015, "Estimating Demand System for US Treasuries," *Working Paper*.

Huberman, G., 2001, "Familiarity Breeds Investment," *Review of Financial Studies*, 14(3), 659–680.

Ivković, Z., and S. Weisbenner, 2005, "Local does as local is: Information content of the geography of individual investors' common stock investments," *The Journal of Finance*, 60(1), 267–306.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp, 2016, "A rational theory of mutual funds' attention allocation," *Econometrica*, 84(2), 571–626.

Kaplan, G., and S. Schulhofer-Wohl, 2012, "Understanding the Long-Run Decline in Interstate Migration," working paper, National Bureau of Economic Research.

Kastl, J., 2016, "Recent Advances in Empirical Analysis of Financial Markets: Industrial Organization Meets Finance," *Working Paper*.

Keloharju, M., S. Knüpfer, and J. Linnainmaa, 2012, "Do Investors Buy What They Know? Product Market Choices and Investment Decisions," 78(6), 2921–2958.

Keloharju, M., S. Knupfer, and E. Rantapuska, 2012, "Mutual fund and share ownership in Finland," *Liiketaloudellinen aikakauskirja*, 2, 178–198.

Kennan, J., and J. R. Walker, 2011, "The Effect of Expected Income on Individual Migration Decisions," *Econometrica*, 79(1), 211–251.

Koijen, R. S., and M. Yogo, 2016, "An Equilibrium Model of Institutional Demand and Asset Prices," *Working Paper*.

Korniotis, G. M., and A. Kumar, 2011, "Do Older Investors Make Better Investment Decisions?," *Review of Economics and Statistics*, 93(1), 244–265.

Lee, C. M., and E. C. So, 2017, "Uncovering Expected Returns: Information in Analyst Coverage Proxies," *Journal of Financial Economics*, 124(2), 331–348.

Lou, D., 2014, "Attracting Investor Attention through Advertising," *Review of Financial Studies*, 27(6), 1797–1829.

Lusardi, A., and O. Mitchell, 2007, "Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education," *Business Economics*, 42(1), 35–44.

Luttmer, E. F. P., 2005, "Neighbors as Negatives: Relative Earnings and Well-Being," *The Quarterly Journal of Economics*, 120(3), 963–1002.

Madsen, J. M., and M. Niessner, 2016, "Is Investor Attention For Sale? The Role of Advertising in Financial Markets," *Working Paper*.

Mankiel, B. G., 1995, "Returns from Investing in Equity Mutual Funds 1971 to 1991," *Journal of Finance*, 50(2), 549–572.

Massa, M., and A. Simonov, 2006, "Hedging, Familiarity and Portfolio choice," *Review of Financial Studies*, 19(2), 633–685.

Meenaghan, T., 2001, "Understanding Sponsorship Effects," *Psychology and Marketing*, 18(2), 95–122.

Merton, R. C., 1987, "A Simple Model of Capital Market Equilibrium with Incomplete Information," *Journal of Finance*, 42(3), 483–510.

Novy-Marx, R., 2013, "The Other Side of Value: The Gross Profitability Premium," *Journal of Financial Economics*, 108(1), 1–28.

Nunn, N., and N. Qian, 2014, "US Food Aid and Civil Conflict," *American Economic Review*, 104(6), 1630–1666.

Odean, T., 1999, "Do Investors Trade Too Much?," *American Economic Review*, 89(5), 1279–1298.

Olson, E. L., and H. M. Thjømøe, 2009, "Sponsorship Effect Metric: Assessing the Financial Value of Sponsoring by Comparisons to Television Advertising," *Journal of the Academy of Marketing Science*, 37(4), 504.

Ortalo-Magné, F., and A. Prat, 2010, "Spatial Asset Pricing: A First Step," *LSE STICERD Research Paper No. TE546*.

Peng, L., and W. Xiong, 2006, "Investor Attention, Overconfidence and Category Learning," *Journal of Financial Economics*, 80(3), 563–602.

Petersen, M. A., 2009, "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies*, 22(1), 435–480.

Petrin, A., and K. Train, 2010, "A Control Function Approach to Endogeneity in Consumer Choice Models," *Journal of Marketing Research*, 47(1), 3–13.

Pirinsky, C., and Q. Wang, 2006, "Does Corporate Headquarters Location Matter for Stock Returns?," *Journal of Finance*, 61(4), 1991–2015.

Puhani, P. A., 2000, "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14(1), 53–68.

Seasholes, M. S., and N. Zhu, 2010, "Individual Investors and Local Bias," *Journal of Finance*, 65(5), 1987–2010.

Shapiro, B., 2017, "Advertising in Health Insurance Markets," *Working Paper*.

Sharpe, W. F., 1964, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, 19(3), 425–442.

Shiller, R. J., and J. Pound, 1989, "Survey Evidence on Diffusion of Interest and Information among Investors," *Journal of Economic Behavior & Organization*, 12(1), 47–66.

Shumway, T., M. B. Szefler, and K. Yuan, 2011, "The Information Content of Revealed Beliefs in Portfolio Holdings," *Working Paper*.

Sims, C. A., 2003, "Implications of Rational Inattention," *Journal of Monetary Economics*, 50(3), 665–690.

Speck, P. S., and M. T. Elliott, 1997, "Predictors of Advertising Avoidance in Print and Broadcast Media," *Journal of Advertising*, 26(3), 61–76.

Tuzel, S., and M. B. Zhang, 2017, "Local Risk, Local Factors, and Asset Prices," *Journal of Finance*, 72(1), 325–370.

Ukman, L., 2017, *IEG's Complete Guide to Sponsorship: Everything You Need to Know About Sports, Arts, Event, Entertainment, and Cause Marketing*. IEG, LLC.

Van Nieuwerburgh, S., and L. Veldkamp, 2009, "Information Immobility and the Home Bias Puzzle," *Journal of Finance*, 64(3), 1187–1215.

Wang, J., 1993, "A Model of Intertemporal Asset Prices under Asymmetric Information," *Review of Economic Studies*, 60(2), 249–282.

Wermers, R., 1999, "Mutual Fund Herding and the Impact on Stock Prices," *Journal of Finance*, 52(2), 581–622.

Wooldridge, J. M., 2010, "Econometric Analysis of Cross Section and Panel Data," Cambridge, MA: MIT Press.

Wooldridge, J. M., 2014, "Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 182(1), 226–234.

Zhu, N., 2002, "The Local Bias of Individual Investors," *Yale ICF working paper*.