U.S. Securities and Exchange Commission
Office of the Investor Advocate

OIAD Working Paper 2022-01

April 2022

How do Consumers Understand Investment Quality?

The Role of Performance Benchmarks

ALYCIA CHIN, Senior Financial Economist, Office of the Investor Advocate, Securities and Exchange Commission, 100 F St., NE, Washington, DC 20549, readlinga@sec.gov, ORCID: 0000-0002-9570-0549

JONATHAN COOK, Financial Economist, Office of the Investor Advocate, Securities and Exchange Commission, 100 F St., NE, Washington, DC 20549, cookjon@sec.gov, ORCID ID: 0000-0001-6067-0960

JAY DHAR, Financial Economist, Office of the Investor Advocate, Securities and Exchange Commission, 100 F St., NE, Washington, DC 20549, dharj@sec.gov

STEVEN B. NASH, Data Analyst, NORC at the University of Chicago, 4350 East-West Hwy 8th Floor, Bethesda, MD 20814, nashst@sec.gov

BRIAN SCHOLL, Chief Economist, Office of the Investor Advocate, Securities and Exchange Commission, 100 F St., NE, Washington, DC 20549, schollb@sec.gov, ORCID ID: 0000-0001-5088-6952

ABSTRACT

We study the impact of mutual fund performance benchmarks on investor decision-making and potential for strategic behavior by firms in displaying benchmarks. In displaying performance, fund companies are required to present a broad-based securities market index ("broad benchmark"), and an optional secondary ("narrow") benchmark that, in some instances, can be more representative of the fund's sector or strategy. Importantly, fund companies have discretion over the choice of benchmarks, within the confines that the benchmarks they select must meet the criteria that the federal securities laws require, presenting opportunity for strategic selection. Our research examines market data and the results of a large behavioral experiment to understand how fund companies employ benchmarks and how investors respond to the presentation of benchmarks.

Standard economic theory does not provide a straightforward role for how benchmarks affect investor decisions. In the experiment, we examine two primary outcomes: (1) subjective attractiveness ratings for a synthetic fund and (2) an incentive-compatible participation outcome that offers participants the choice between our fund and a guaranteed return over a six-month holding period. We administer treatment conditions that vary the number of benchmarks presented, the relative position of the benchmark vis-à-vis our synthetic fund, the use of broad or narrow benchmarks, and the use of narrative text. Our results indicate that investors respond to benchmarks. In particular, subjective attractiveness ratings are much lower when participants view fund performance accompanied by a single benchmark that outperforms the fund. This decrease in attractiveness also occurs, to a lesser extent, when participants view two benchmarks that both outperform and underperform the fund. Allocations to the synthetic fund are also lower when participants see a single benchmark above the fund. Surprisingly, participants with higher investment sophistication appear to react most strongly to benchmarks (rather than lower sophistication individuals). Additionally, the distinction between narrow and broad benchmarks and the narrative descriptive text about the benchmarks do not have a differential impact beyond the position of the benchmark. Finally, using an economic model, we ask what type of benchmark presentation gets investors closest to their optimal allocation, finding that conditions with no benchmark and with two benchmarks minimize distortions.

Using data from the Morningstar Direct database, we contextualize these findings and the concerns that our results raise in situations where funds have discretion regarding the selection of benchmarks. Specifically, we document performance variation of benchmarks within a given sector, as well as the decision to present a secondary benchmark. Ultimately, these patterns raise the possibility that funds can pick benchmarks that satisfy the requirements for permissible benchmarks, but are relatively poor performing as compared to other permissible benchmarks. This would put the fund's relative performance in a more positive light, which may affect investors' evaluations and investment decisions.

ACKNOWLEDGEMENTS

We thank staff at the RAND Corporation and Ipsos for helping us conduct this research, including Andrew Parker, Katie Carman, Vanessa Parks, and Ying Wang. We also thank staff of the Securities and Exchange Commission, particularly members of the Office of the Investor Advocate and Division of Investment Management, for helpful comments on our research design and background on legal requirements. Brianna Middlewood contributed significant experimental design and survey expertise during her tenure at the SEC. Steven Nash is an onsite institutional contractor to OIAD. Rick Fleming provided excellent mentorship in the role of the SEC's first Investor Advocate.

Contents

Abstract

We study the impact of mutual fund performance benchmarks on investor decision-making and potential for strategic behavior by firms in displaying benchmarks. In displaying performance, fund companies are required to present a broad-based securities market index ("broad benchmark"), and an optional secondary ("narrow") benchmark that, in some instances, can be more representative of the fund's sector or strategy. Importantly, fund companies have discretion over the choice of benchmarks, within the confines that the benchmarks they select must meet the criteria that the federal securities laws require, presenting opportunity for strategic selection. Our research examines market data and the results of a large behavioral experiment to understand how fund companies employ benchmarks and how investors respond to the presentation of benchmarks.

Standard economic theory does not provide a straightforward role for how benchmarks affect investor decisions. In the experiment, we examine two primary outcomes: (1) subjective attractiveness ratings for a synthetic fund and (2) an incentive-compatible participation outcome that offers participants the choice between our fund and a guaranteed return over a six-month holding period. We administer treatment conditions that vary the number of benchmarks presented, the relative position of the benchmark vis-à-vis our synthetic fund, the use of broad or narrow benchmarks, and the use of narrative text. Our results indicate that investors respond to benchmarks. In particular, subjective attractiveness ratings are much lower when participants view fund performance accompanied by a single benchmark that outperforms the fund. This decrease in attractiveness also occurs, to a lesser extent, when participants view two benchmarks that both outperform and underperform the fund. Allocations to the synthetic fund are also lower when participants see a single benchmark above the fund. Surprisingly, participants with higher investment sophistication appear to react most strongly to benchmarks (rather than lower sophistication individuals). Additionally, the distinction between narrow and broad benchmarks and the narrative descriptive text about the benchmarks do not have a differential impact beyond the position of the benchmark. Finally, using an economic model, we ask what type of benchmark presentation gets investors closest to their optimal allocation, finding that conditions with no benchmark and with two benchmarks minimize distortions.

Using data from the Morningstar Direct database, we contextualize these findings and the concerns that our results raise in situations where funds have discretion regarding the selection of benchmarks. Specifically, we document performance variation of benchmarks within a given sector, as well as the decision to present a secondary benchmark. Ultimately, these patterns raise the possibility that funds can pick benchmarks that satisfy the requirements for permissible benchmarks, but are relatively poor performing as compared to other permissible benchmarks. This would put the fund's relative performance in a more positive light, which may affect investors' evaluations and investment decisions.

Keywords: mutual fund performance, benchmarks

1. Introduction

Every day, American investors use a variety of financial products to pursue their financial goals. Investors express interest in using mutual funds[1] to fund retirement, save for educational expenses, and protect against emergencies (ICI, 2021a), contributing to a growing, $25 trillion mutual fund industry.

To ensure that investors receive the information they need to make decisions about investments, regulations require financial institutions to provide "disclosures," informational documents that include product terms and agreements (Kozup et al., 2012). Numerous regulations require disclosures of important attributes of investment products. For example, financial regulations require that disclosures such as the "prospectus" document contain a wealth of information on fees and expenses, risks, objectives, and performance (for requirements for open-ended funds, see Form N-1A, the registration form for these funds). Despite the prevalence of disclosure requirements, there is significant debate about the ability of consumers to comprehend mandatory disclosures and the corresponding usefulness of these disclosures to guide decisions (e.g., Ben-Shahar and Schneider, 2011).

The current research examines industry practices regarding historical performance information and disclosures of that information. In particular, we examine fund choices of mutual fund "benchmarks," comparisons that are required to be present in many fund disclosures, and that may help investors contextualize fund performance; the requirements for benchmarks are described further in Section 2.

We concentrate on performance for a few reasons. First, this is an area that is important to existing and prospective investors; investors report that performance information is important to them (ICI, 2021b) and significant research, described further below, shows that performance information attracts attention. Second, the normative and descriptive roles of benchmark information in decision-making are not entirely clear, with different theories providing different guidance regarding whether benchmarks should be impactful or ignored. Potential disagreement about the role of benchmarks makes this area a fruitful one for empirical testing. Finally, a recent rule proposal by the Securities and Exchange Commission addressed, in part, funds' use of performance benchmark indexes.[2]

In this paper, we use several research methods to triangulate the role and the effects of benchmarks. Most importantly, we conducted a behavioral experiment using a large, nationally representative study population to determine how investors' evaluations of funds and investment behavior respond to benchmarks, and we conducted extensive market data analysis to understand how benchmarks are used. We also conducted a small number of formative qualitative interviews with investors to better design our main research methodologies.

[1] In this paper, the term "funds" refers to open-end funds registered on Form N-1A.

[2] *See* Tailored Shareholder Reports, Treatment of Annual Prospectus Updates for Existing Investors, and Improved Fee and Risk Disclosure for Mutual Funds and Exchange-Traded Funds; Fee Information in Investment Company Advertisements, Investment Company Act Release No. 33963 (Aug. 5, 2020) [85 FR 70716 (Nov. 5, 2020)]. ("2020 Shareholder Reports Proposal").

1.1 Mutual Fund Performance Disclosures

Extant research demonstrates that investors care about historical performance of investments. Research consistently shows that investors prioritize information on investment performance (Barber, Odean, and Zheng 2005; Pontari, Stanaland, and Smythe 2009; Scholl, Craig, and Chin, 2022). One common theory for why investors weigh historical information heavily is that they expect historical returns to persist. Indeed, attention to performance information persists even in the face of statements that funds are required to include in their disclosure that the fund's past performance is not necessarily an indication of how the fund will perform in the future (Johnson, Tellis, and VanBergen, 2022).[3]

When presenting historical performance data in prospectuses and shareholder reports, funds are required to provide a benchmark that investors can use to make comparisons. A fund references an "appropriate broad-based securities market index," which we refer to as a "broad benchmark" for brevity. These benchmarks represent broad market activity (e.g., S&P 500). Funds may also reference additional, more narrowly based indexes that reflect the market sectors in which the fund invests, which we refer to as "narrow benchmarks" (more details in the next section). For instance, a fund specializing in the materials sector might display its performance against a materials sector index (a "narrow" benchmark).

There are at least three theories regarding why benchmarks could affect investors' decision-making. First, if investors are imperfectly informed about the distribution of performance information – possibly because it is difficult to search through an industry with over 8,000 mutual fund options – then providing a benchmark could provide a shortcut to distributional information that allows investors to avert costly search (Hortaçsu and Syverson, 2004). Second, a benchmark could provide information about market shocks ("factors" in arbitrage pricing theory), contextualizing factors and events that the fund cannot avoid. A narrow benchmark provides information about the return relative to the factors that the fund is exposed to. Again, following this theory, benchmark information could provide information about the overall performance of a fund. Third, psychological theory suggests that, to increase understanding and help people with unfamiliar or otherwise difficult-to-evaluate products, disclosures should provide decision makers with meaningful comparisons (e.g., Chin and Bruine de Bruin 2019; Hsee 1996; Hsee and Zhang 2010; Larrick et al. 2015). As such, it is possible that benchmarks help drive evaluations by providing a salient comparison. When the Commission adopted the requirement to present fund performance against an appropriate broad-based securities market index, the Commission stated that the index comparison requirement is designed to show how much value the management of the fund added by showing whether the fund "out-performed" or "under-performed" the market.[4]

---

[3] *See, e.g.*, Items 4(b)(i) and 27(b)(7)(ii) of Form N-1A; rule 482(b)(3)(i) under the Securities Act of 1933.

[4] *See* Disclosure of Mutual Fund Performance and Portfolio Managers, Investment Company Act Release No. 19382 (Apr. 6, 1993) [58 FR 19050 (Apr. 12, 1993)] ("1993 Mutual Fund Performance Disclosure Final Rules"); *see also* Tailored Shareholder Reports, Treatment of Annual Prospectus Updates for Existing Investors, and Improved Fee and Risk Disclosure for Mutual Funds and Exchange-Traded Funds; Fee Information in Investment Company Advertisements, Investment Company Act Release No. 33963 (Aug. 5, 2020) [85 FR 70716 (Nov. 5, 2020)]

There are also reasons why benchmarks could have limited effects. In situations where a fund's historical performance is disclosed, and the fund's performance itself is the decision-relevant attribute, it is not clear what information is gained from a benchmark. The strictest reading of a classic Rational Expectations framework, which assumes perfect information and no limitations on information processing ability, would imply that investors would be highly informed regardless of the fund's provision of a benchmark. While these assumptions may not be tenable for real-world investment behavior, especially for retail investors, it can be useful to treat this framework as a logical comparison. A second reason why benchmarks may have a limited impact is that some investors may believe that funds choose benchmarks strategically, in an attempt to influence investor evaluations. In this case, investors may consciously attempt to ignore benchmark comparisons. Third, investors who do not understand what the benchmark information is supposed to represent may ignore it. Thus, there are some reasons why investors may not respond to benchmark information.

Finally, if a benchmark is not well-matched to a fund, it could provide a confusing or distorting signal about whether a fund is performing relatively well. Active share is defined as the percentage of a fund's holdings that differ from their benchmark (Cremers and Petajisto, 2009); funds that have higher active share have more potential to deviate from their benchmarks. Indeed, this critique has been raised by industry in stating that, for instance, specialized sector funds should not need to be compared to a "broad-based" benchmark, like the S&P 500 (Fidelity, 2021; ICI, 2020; John Hancock, 2021). If investors face benchmarks they believe are not well-matched, it is possible they find that information irrelevant. At the same time, financial regulations offer funds at least some discretion on the choice of benchmarks.[5] This discretion raises the possibility that some funds could choose benchmarks strategically to make the fund appear more attractive to current or potential investors. Prior work, focusing on the role of narrow benchmarks, has found that some funds' benchmarks do not provide the best match in terms of exposure to market factors (as in Sensoy, 2009) or in terms of holdings (as in Cremers, Fulkerson, and Riley, 2022). The extent to which such strategic selections occur, and the extent of their influence on investors, remains an open question for future research. Evaluating funds relative to their benchmarks can also give rise to other behaviors. There are incentives for fund managers to incorporate their benchmark in their fund's holdings to hedge against poor performance relative to the benchmark (Pavlova and Sikorskaya, 2022).

Proposal (proposing changes to funds' shareholder report contents and presentation, but proposing to retain the requirement for funds to present performance in relation to an appropriate broad-based securities market index).

[5] *See* Instruction 5 to Form N-1A Item 27A(b)(7) (defining "appropriate broad-based securities market index") and Instruction 6 to Form N-1A Item 27A(b)(7) (encouraging a fund, in addition to comparing its performance to the required broad-based index, also to include other more narrowly based indexes that reflect the market sectors in which the fund invests). Both instructions provide flexibility to the fund to choose the indexes it includes in its performance presentation, within the parameters that the instructions specify. *See also* Disclosure of Mutual Fund Performance and Portfolio Managers, Investment Company Act Release No. 19382 (Apr. 6, 1993) [58 FR 19050 (Apr. 12, 1993)] (stating that the final rules' instruction requiring the inclusion of an appropriate broad-based securities market index "gives a fund considerable flexibility in selecting a broad-based index that it believes best reflects the market(s) in which it invests)."

1.2 Related Literature

Our work is related to several existing areas of academic research. There are two closely related papers. The first is Sensoy (2009), which finds that mutual funds flows respond to the performance of the fund relative to the prospectus benchmark. The second is Mullaly and Rossi (2022), which analyzes changes to mutual funds' self-declared benchmarks using prospectus data. This paper finds that funds change indexes in a manner that improves relative-benchmark performance; that is, they are more likely to add indexes with lower past returns and drop indexes with higher past returns. Unlike these papers, we do not analyze benchmark changes. Instead, we examine the mechanism behind investors' decisions, including perceptions of future risk and return and performance relative to other options. Additionally, we use a mix of experimental and industry data, whereas these authors concentrate on fund data.

We also contribute to several broader literatures. First, a large and growing literature conducts randomized evaluations of information provision. For mutual funds in particular, related papers include Choi, Laibson, and Madrian (2010); Kozup, Howlett, and Pagano (2008); and Thorp, Bateman, Dobrescu, Newell, and Ortmann (2020). Within household finance, similar work is conducted by Chin and Bruine de Bruin (2019) for credit cards, Lacko and Pappalardo (2010) for mortgages, and Chin et al. (2022) for overdraft.

Second, we contribute to literature on households' subjective probabilities (for a review, see Bruine de Bruin, Chin, Dominitz, and van der Klauuw, 2022) and more specifically, how information experiments affect beliefs. The number of papers on this topic are growing, including for topics like inflation and home prices (e.g., Armantier, Nelson, Topa, van der Klaauw and Zafar, 2016; Armona, Fuster, and Zafar, 2016).

Third, we speak to research examining search costs within the investment industry. Various papers model retail investors as having high search costs, assuming that investors randomly sample other mutual funds and stop when search costs are "too high" (see Hortaçsu and Syverson, 2004). Survey data from Choi and Robertson (2020) support the idea of search costs, as 40% of non-investors in their nationally representative sample say that the costs of learning about stocks are an important factor in why they do not participate in the stock market. Other empirical papers include Roussanov, Ruan, and Wei (2021), in which an average investor implicitly incurs a cost equivalent of foregoing 0.39% return on investment every time a fund is sampled. Hortaçsu and Syverson (2004) estimate search costs for index funds between 11 and 20 basis points.

1.3 Research Overview

We proceed in the following sections: First, we describe the institutional background in detail, including regulatory requirements on fund disclosure of benchmarks (Section 2). Next, we describe the state of benchmark disclosure using an analysis of the Morningstar Direct database, which includes data on mutual funds and their associated benchmarks. We provide statistics on the prevalence of certain common benchmarks and show that, within every sector, there are at

least a dozen unique benchmarks for funds to choose from (Section 2.1). Performance of these benchmarks can vary by over 400% over a 10-year period (Section 2.2). In Sections 3 and 5, we describe the setup and results of our experiment, in which we measure how investors and non-investors respond to variation in disclosure of mutual fund benchmarks. We find that participants respond to benchmark presentations, with variation in subjective evaluations of funds and incentivized investment decisions. Perhaps surprisingly, given the prominence of beliefs regarding future performance in economic models of investing (e.g., Markowitz, 1952; Sharpe, 1964), and the role expectations often play in theoretical and empirical work on investor behavior (e.g. Giglio, Maggiori, Stroebel, and Utkus, 2021; Barberis, Jin, and Wang, 2021), we find limited evidence that expectations of future performance differ across conditions (Section 3.5). Also surprisingly, non-investors – the least sophisticated participants – were relatively unaffected by benchmark presentation (Section 3.4). In Section 6, we describe survey results regarding benchmarks. In Section 7, we return to an analysis of Morningstar Direct to provide further descriptive evidence on the potential for strategic behavior by funds in benchmark disclosure. Section 8 summarizes and concludes.

2. Institutional Background on Benchmark Requirements

Financial regulations require funds to provide comparative information when presenting performance data in their shareholder reports.[6] Specifically, if sufficient history is available, funds must provide a line graph that shows 10 fiscal years of performance, accompanied by an "appropriate broad-based securities market index," in annual shareholder reports that are provided to existing investors. Funds may also provide this line graph in semi-annual shareholder reports. In both reports' line graphs, funds have the *option* of presenting performance relative to one or more additional indexes. These additional indexes can be "broad-based," as with the first, or more narrowly tailored to the assets and strategy of the fund.[7] For brevity, we refer to both of these indexes as "benchmarks," and distinguish between "broad-based" and "narrow" benchmarks. Benchmarks also are required to be provided in the performance disclosure that appears in funds' statutory prospectuses and summary prospectuses, and they commonly are provided in fund advertising as well.

[6] *See id.*

[7] Mutual funds' prospectus and shareholder report disclosures are governed by Form N-1A. Instruction 5 to Item 27, "Financial Statements" in Form N-1A states, "For purposes of this Item, an "appropriate broad-based securities market index" is one that is administered by an organization that is not an affiliated person of the Fund, its investment adviser, or principal underwriter, unless the index is widely recognized and used. Adjust the index to reflect the reinvestment of dividends on securities in the index, but do not reflect the expenses of the Fund." Instruction 6 to this Item states, "A Fund is encouraged to compare its performance not only to the required broad-based index, but also to other more narrowly based indexes that reflect the market sectors in which the Fund invests. A Fund also may compare its performance to an additional broad-based index, or to a non-securities index (e.g., the Consumer Price Index), so long as the comparison is not misleading." See https://www.sec.gov/files/formn-1a.pdf.

2.1 Distribution of Benchmarks

The benchmarks chosen by mutual funds that appear in their statutory prospectuses are captured in the Morningstar Direct open-ended fund database. We analyze data on benchmarks from Morningstar database as of March 2022, the most recent complete month of data available as of the time of this writing. In these analyses, we concentrate only on equity funds. We remove target date funds, since they are less likely to concentrate on performance than changes in risk profile; this focus is reflected in use of blended benchmarks with weights that are continually adjusted as they approach their target date. We also remove index funds because their goal is to track benchmarks. This brings our data set to 3,187 mutual funds. Benchmarks are the same for all share classes within a given fund.

The Morningstar database captures "primary" and "secondary" indexes. However, because there is no ordering requirement for fund benchmarks, Morningstar's identification of primary and secondary benchmarks could provide an imperfect mapping to "broad" and "narrow" indexes as described in regulatory requirements. As such, we reclassified benchmarks as broad and narrow based on the correlation of the benchmark with the S&P 500, so that the benchmark with the highest correlation was identified as the broad benchmark and subsequent benchmarks were considered as the secondary. In the text, we occasionally use "broad" and "primary" and "narrow" and "secondary" interchangeably.[8]

Nearly all funds have at least one benchmark listed, with a handful of missing benchmarks in the data for funds with recent inception dates.[9] In contrast, approximately 70% of funds choose not to include a second benchmark (left panel of Figure 1). Table 1 displays the 10 most common primary and secondary benchmarks in the database. In cases where a fund has two benchmarks listed, we define the primary benchmark as the benchmark with the largest correlation with the S&P 500 Index,[10] which is the most commonly used benchmark. As shown in the table, the S&P 500 Total Return Index accounts for 23% of the primary benchmarks. Among the list of the most common benchmarks, there are a set of arguably broad benchmarks, a set of arguably sector-specific benchmarks, and a set of global or emerging market benchmarks. Some indexes appear in both lists (e.g., Russell 2000 Value Total Return and Russell 1000 Growth Total Return). The table does not show the least common benchmarks. In the top 12

[8] In considering individual fund's choices of benchmarks, we also found cases in which the benchmarks chosen by funds were difficult to interpret in the sense of broad and narrow benchmarks. In separate analysis (not shown) we also used alternative definitions corresponding to "broad" and "narrow", including Morningstar classifications of Primary and Secondary. Analyses from these classifications provided qualitatively identical and quantitatively similar results. Note also that Morningstar captures only two benchmarks (or less) for each fund, but we have observed cases in which more than two benchmarks are used by a fund.

[9] To verify the data, we randomly selected 105 funds and pulled benchmark information for those funds from the funds' prospectus documents. We found that the primary benchmarks matched in 104 cases and did not match in one case. The secondary benchmarks matched in 97 cases, and did not match in eight cases.

[10] According to the S&P Dow Jones Indices website (https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview), "The S&P 500® is widely regarded as the best single gauge of large-cap U.S. equities. According to our Annual Survey of Assets, an estimated USD 13.5 trillion is indexed or benchmarked to the index, with indexed assets comprising approximately USD 5.4 trillion of this total (as of Dec. 31, 2020). The index includes 500 leading companies and covers approximately 80% of available market capitalization." In addition to being the most commonly used benchmark by fund companies, it is commonly used in academic studies, and it is widely recognized: "in the US, the most widely known market value-weighted stock index is the S&P 500" (Beneish and Whaley, 1997); and "The S&P 500 Index is widely recognized as reflecting the overall state of the U.S. economy…" (Latham and Braun, 2010).

sectors by fund count, 10.8% of funds use a primary benchmark used by less than 5 funds in their sector. For secondary benchmarks, this figure is 16.0%.

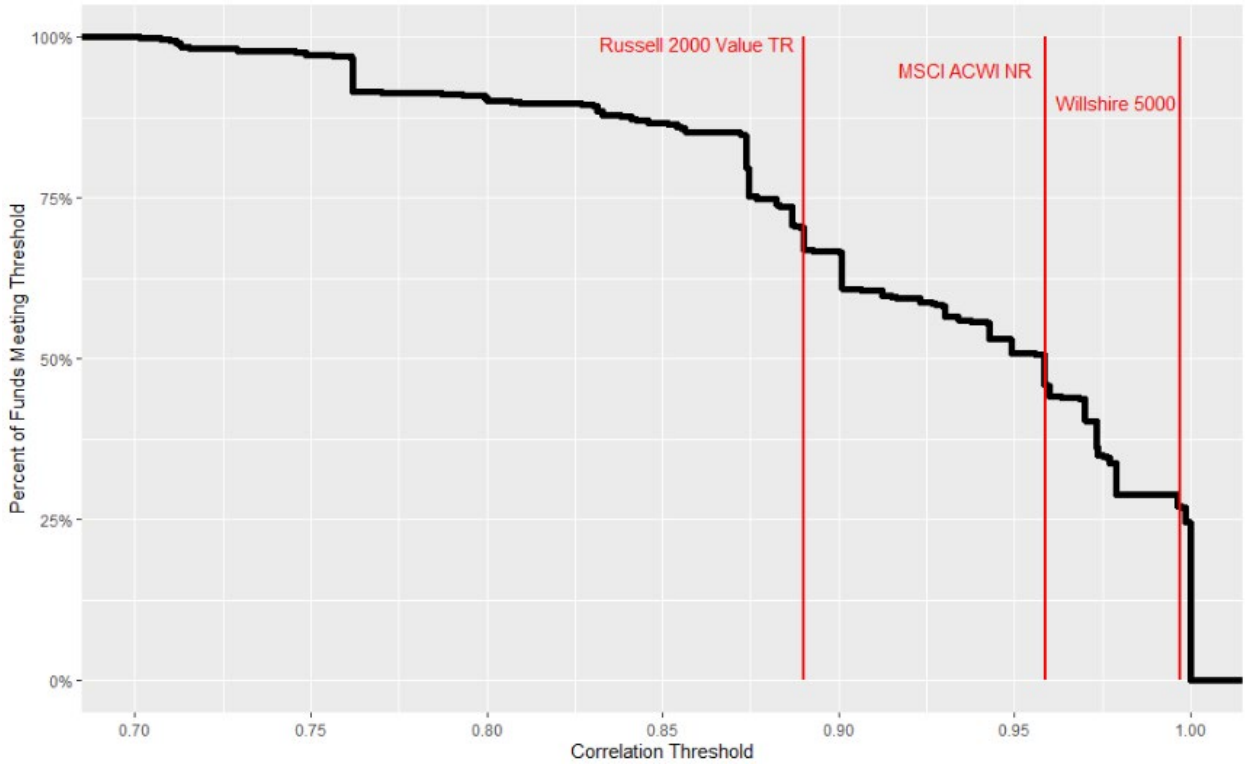
Figure 1. Number of Unique Benchmarks and Percentage of Funds with Benchmarks.

Primary Benchmark	n funds	Percent	Secondary Benchmark	n funds	Percent
S&P 500 TR USD	686	23.21	Russell 1000 Value TR USD	39	4.13
Russell 2000 TR USD	153	5.18	Russell 1000 Growth TR USD	35	3.71
MSCI EM NR USD	146	4.94	Russell 2000 Value TR USD	29	3.07
Russell 1000 Value TR USD	141	4.77	Russell 2000 Growth TR USD	24	2.54
MSCI EAFE NR USD	138	4.67	Russell 2000 TR USD	17	1.80
MSCI ACWI Ex USA NR USD	125	4.23	Russell Mid Cap Growth TR USD	16	1.69
MSCI ACWI NR USD	120	4.06	DJ US Total Stock Market Float Adj USD	14	1.48
Russell 1000 Growth TR USD	117	3.96	MSCI EAFE NR USD	14	1.48
Russell 2000 Value TR USD	96	3.25	MSCI ACWI NR USD	12	1.27
MSCI World NR USD	91	3.08	Russell 1000 TR USD	11	1.17

To better understand the relationship between these indexes, we next explored the correlations between them. Specifically, we calculated the correlation between each primary index and the S&P 500 Index using monthly data over the past 10 years. Figure 2 displays the

proportion of funds whose benchmark meets or exceeds a given correlation "threshold." The curve is downward sloping, demonstrating that, as the correlation threshold increases (to the right on the graph), the proportion of funds meeting that threshold necessarily decreases. The 23% of funds that use the S&P 500 Index as their primary benchmark are displayed at the right-most extreme of the graph, with a correlation of 1.00. Finally, the red vertical lines display example correlations between a selected index and the S&P 500 Index. As shown, other broad-based security market indexes (e.g. the Wilshire 5000 Index) were extremely highly correlated with the S&P 500 Index. An index like the MSCI ACWI Index, which reflects large- and mid-cap stocks,[11] has a correlation of 0.96. Notably, many of the most common sector benchmarks also were highly correlated with the S&P 500 Index (about 0.90 to 0.97). Even among the common global indices, some indices had a correlation with the S&P 500 Index of over 0.95. Nevertheless, in our data, only about half of funds present at least one benchmark that has a correlation with the S&P 500 Index of 0.95 or above (Figure 2).

Figure 2. Distribution of Benchmark Correlations with S&P 500 Index.



[11] According to the MSCI website (https://www.msci.com/our-solutions/indexes/acwi), "The MSCI ACWI Index, MSCI's flagship global equity index, is designed to represent performance of the full opportunity set of large- and mid-cap stocks across 23 developed and 24 emerging markets. As of May 2022, it covers more than 2,933 constituents across 11 sectors and approximately 85% of the free float-adjusted market capitalization in each market."

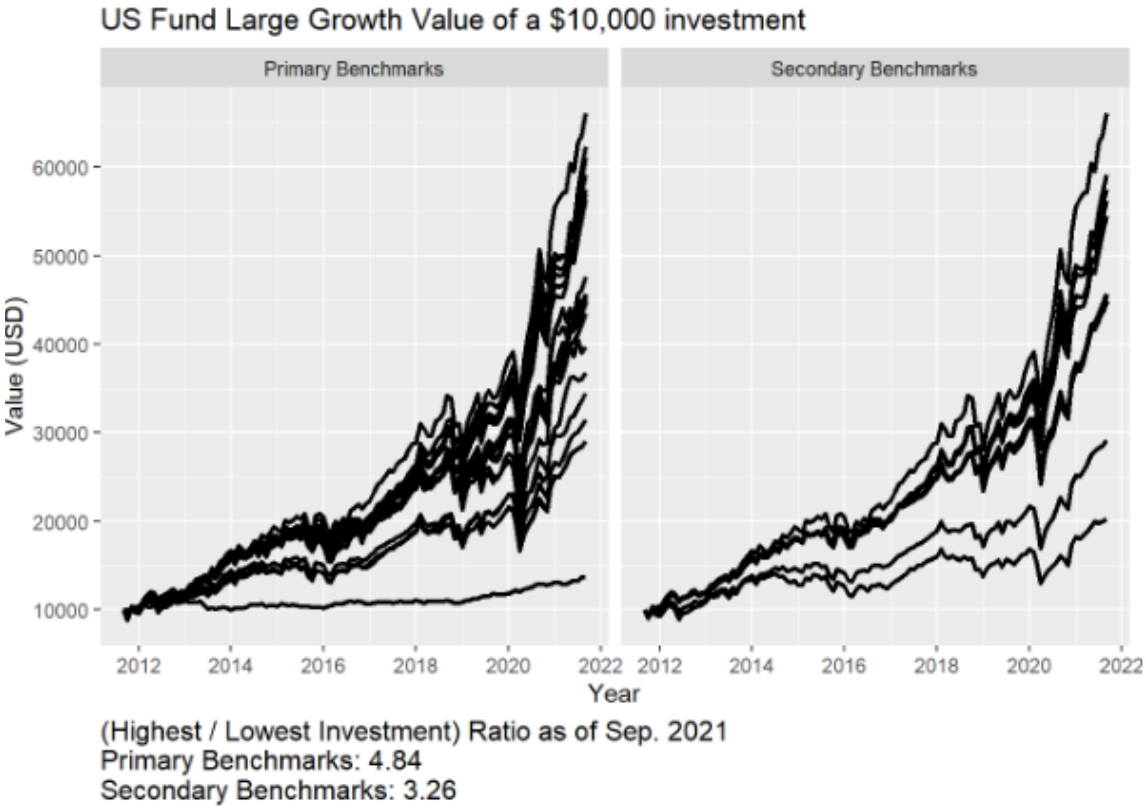
Note. This figure displays the proportion of funds that have a benchmark with the S&P 500 Index at a given correlation threshold or higher. Correlations are calculated using monthly levels of the benchmark and S&P 500 Index.

2.2 Performance Variation in Benchmarks

Figure 1 also shows number of unique benchmarks used for each of the top 12 sectors (by fund count). It suggests that, at the minimum, funds in a given sector use at least 10 distinct benchmarks. However, the number of benchmarks that exist within a given sector can be higher. For example, funds in the Large Blend category jointly use 39 unique secondary benchmarks.

To further explore variation in of benchmark performance, Figure 3 below illustrates the degree of variation in the performance of benchmarks used. Specifically, we plot the value of a hypothetical $10,000 investment over the period September 2011 to September 2021 for each benchmark in each sector (see additional sectors in Appendix A). The charts are split by primary and secondary benchmarks. In every case, there is a wide gap between the returns of the best and worst performing benchmarks; in Figure 3, for example the ratio of highest to lowest return benchmarks is 4.84 for primary benchmarks, meaning that performance could be 484% higher for a high-performing benchmark versus a low-performing one within the same sector. This variation demonstrates that funds within a given sector are using benchmarks that vary widely in performance.

Figure 3. Variation in benchmark performance among US Large Growth Value Funds.



Note. Graph displays performance of a $10,000 investment over a 10-year period. Each black line represents an index that is used by a fund. See Appendix A for additional examples.

3. Experimental Design

To understand the role of benchmarks in investors' decision-making, we use a controlled experiment that helps abstract away from the specific properties of any particular benchmark. In the experiment, we vary displays of cumulative performance information in a between-subjects design, and measure participants' subjective evaluations and investment choices. The experimental design consists of eight randomly assigned benchmark presentation conditions. We focus on two main outcomes: our primary outcome of interest is participants' subjective evaluation of the attractiveness of a hypothetical "Middlewood Materials Fund"; our secondary outcome relates to an incentive-compatible participation game in which participants chose to allocate a fraction of a hypothetical $15,000 investment balance between investment in the fund and a savings account paying a fixed interest rate. Additionally, to better understand consumers' beliefs about benchmarks, we collect nationally representative survey data on different statements about benchmarks.

3.1 Qualitative and Quantitative Pilot Studies

Before running our experiment, we conducted qualitative and quantitative pilot studies. Our qualitative pilot study included interviews with 16 geographically disbursed U.S. investors, recruited from the AmeriSpeak panel administered by NORC at the University of Chicago. All of the participants reported owning "mutual funds, exchange traded funds (ETFs) or similar pooled investments" in a screening survey. Additionally, we sampled participants with a range of both high and low mutual fund literacy (as assessed by Scholl and Fontes, 2021). Interviews took place online in October and November 2021.

During the interviews, participants reviewed a three-page mockup of a shareholder report that featured a hypothetical "Middlewood Small Cap Fund." The interviews started by collecting general impressions of the document and understanding of the fund's fees. Next, participants viewed four versions of a 10-year performance graph.[12] For all participants, presentation of graphs went as follows: the first graph showed the fund's performance alone; the second was a randomly assigned graph that displayed the fund with either a narrow benchmark or broad-based benchmark (represented by the Russell 2000 Small Cap Value Index or the S&P 500 Index, respectively); a third graph displayed the fund with both benchmarks; and the final graph included text explaining the benchmarks (see Appendix B). The rationale for introducing benchmarks in this way was to gain initial insight on how introduction of benchmark information could affect fund evaluations after the participant had provided an initial impression without the benchmark. As noted in Section 2 above, the S&P 500 Index is the most common broad-based market index. The Russell 2000 Small Cap Value Index was selected by examining the performance of all narrow benchmarks currently used by small cap funds with at least ten years of performance history in their prospectus disclosures and selecting the benchmark representing US small cap funds with the worst cumulative performance over the prior ten years. This allowed us to gain preliminary insight on how investors might react to a benchmark line that is relatively poor performing over the period. In the qualitative study, the monthly returns of the Middlewood Small Cap Fund were generated by adding a small positive alpha and some noise to the narrow benchmark. The noise was generated such that for a random 20% of months a small amount was added or subtracted from the returns. This was done so that the fluctuations of the fund and narrow benchmark did not match exactly, but so that fund volatility did not differ substantially from the benchmark.

For the purposes of the current research, we highlight only three findings from the interviews (see some additional discussion in Appendix B). First, they suggested that benchmarks could affect participants' interpretation of mutual fund performance, as all participants stated something about relative performance between the fund and one or both benchmarks. For example, one participant noted, "Clearly the fund has outperformed the small cap value index, fairly significantly over time" (Male, 65 years old). Most participants appeared

[12] Graphical performance information is often contained in other informational content such as fund prospectuses, as well as fact sheets and other advertisements. These other informational sources have differing regulations on the presentation of information. While our findings in this study are generalizable to the use of graphical benchmark information in many contexts, we focused on the requirements for shareholder reports.

to react to the inclusion of benchmarks by updating their subjective evaluation of the fund in reaction to the relative position of the line. So when a reference line was provided with inferior performance over the ten-year period, we tended to observe that participants updated their evaluation of the fund in a more positive way, while a reference line with superior performance seemed to lead to a more negative impression of the fund. A second preliminary takeaway was that some participants were not familiar with specific benchmarks, for instance, "I don't know what the Russell 2000 is, and I can't compare against something where I don't know what it is" (Female, 24 years old). Such confusion led us to attempt to clarify the benchmarks by adding text describing the benchmarks underneath the graphs (as described in the "narrative" conditions below). Third, participants mentioned some (often mistaken) beliefs in response to the performance graph. For example, one participant stated, referring to the fund and the narrow index, that "One is 'value fund' and one is 'value index' so it's not clear if the index is part of the fund." Following the mental models approach (Morgan, Fischhoff, Bostrom, and Atman, 2001), such statements informed particular survey items we administered in the experiment, primarily described in Section 6.

Our quantitative pilot study was conducted in March 2022 and included 366 participants recruited from the Ipsos Knowledge Panel, which is also used for the full experiment. The main purpose of this pilot was to evaluate the specific framework used for the allocation decisions described below, the appearance of the stimuli on personal devices, the overall length of the survey, and other operational details of survey administration (e.g., sampling). Following the quantitative pilot, we simplified certain question to reduce respondent burden. Pilot participants are not included in the analyses below.

3.2 Stimuli Selection and Construction

For this study, we carefully designed our stimuli (for additional detail, see Appendix C). Our research questions required two criteria be met. First, we needed to present narrow and broad-based benchmarks so that we could determine whether this classification differentially affected participants' reactions to the disclosed information. Second, to isolate the effect of benchmark classification and avoid confounding effects of performance differences, we needed to be able to present narrow and broad-based benchmarks with identical performance. Further, we believed that presenting benchmarks that performed both better than, and worse than, the Middlewood Fund would provide the most interesting theoretical variation.

To satisfy these criteria, we used the Morningstar Direct database to identify a naturally occurring set of four benchmarks. Specifically, we selected two narrow benchmarks that could apply to a materials fund. We also identified two broad-based benchmarks that had similar performance to the two narrow benchmarks. In the end, this process yielded two pairs of benchmarks; in each pair, there was one narrow and one broad-based benchmark that had similar cumulative 10-year performance and variance. Between the two pairs, there was a performance difference (11% vs. 16% annual return on average over 10 years); the Middlewood Materials

Fund is a synthetic fund constructed to fall between these two figures. The fact that we were able to identify such pairings of benchmarks used by actual funds within an actual market sector highlights the flexibility of current disclosure rules (i.e., the discretion that funds have over the choice of benchmarks) and the potential for strategic selection of benchmarks by firms.

To provide the impression that the benchmarks were broad-based or narrow, we named them the "Imprimiis Total US Market 1000 Index" or "Imprimiis Materials Select Index," respectively. In certain "narrative" conditions, we addressed the potential concern raised by participants in the qualitative pretest that they were unfamiliar with certain benchmarks. Specifically, we explained the meaning of the two benchmarks by saying "This graph compares the Middlewood Materials Fund to two indexes. The first index, the Imprimiis Total US Market 1000 Index, allows you to see how the fund is performing relative to the US stock market as a whole. The second index, the Imprimiis Materials Select Index, allows you to see how the fund is performing relative to an index with similar investments in the materials sector."[13] This text was reviewed by securities market experts to ensure it was realistic.

3.3 Recruitment and Sample Characteristics

We recruited participants using the Ipsos Knowledge Panel, a nationally representative internet panel.[14] The Ipsos panel includes approximately 60,000 members who were recruited via probability-based sampling methods. The Knowledge Panel provides computers and internet connections for respondents who do not have them at the time of panel recruitment. Each panelist provides basic demographics upon enrollment, as well as survey responses on various topics (such as financial behaviors, a subjective assessment of their credit score, and health insurance coverage). They receive incentives for completing each survey and are automatically entered into sweepstakes for additional gift cards or cash. For this survey, respondents could also receive payments for their investment decisions, which we describe below. We began with a subsample of the Knowledge Panel that included only US citizens aged 18 or older.

Table 2 presents summary data for the samples of valid observations that were randomized into treatment or control. Demographic characteristics are drawn from data that the respondent provided to Ipsos upon enrollment in the panel. Missing covariates were collected via survey questions at the end of the experiment if necessary. Using a variety of procedures, we verified that random assignment worked insofar as the experiment was balanced (for additional detail on one, see Appendix D).

[13] In another condition, we had an additional sentence saying, "These indexes allow you to better understand the performance of alternative investment options." However, there were no significant differences between this longer narrative and its shorter counterpart on our primary outcome measures, and therefore we collapsed these two conditions in our analyses.

[14] This Ipsos panel has been used by many other studies, including reports by other regulators, see, for example, Consumer Financial Protection Bureau (2020).

Table 2. Summary statistics for study participants.

Variable	Mean	Std. dev.
Age	52.40	16.90
Male	0.51	0.50
Income in $1000s (based from bin midpoints)	95.00	55.70
Net assets ($)[15]	422,000	950,000
Race/ethnicity		
White Non-Hispanic	0.71	0.46
Black Non-Hispanic	0.10	0.30
Other Non-Hispanic	0.05	0.22
Hispanic	0.11	0.32
Two or More Races	0.03	0.17
Education		
No high school diploma or GED	0.06	0.23
High school graduate (high school diploma or the equivalent GED)	0.25	0.43
Some college or Associate's degree	0.28	0.45
Bachelor's degree	0.24	0.43
Master's degree or higher	0.18	0.39
Investor type		
Non-investors	0.32	0.47
Retirement-only investors	0.24	0.43
Independent investors	0.45	0.50
Investment ownership		
Report owning investments that track the overall US stock market, like an S&P 500 Index fund or a Dow Jones Industrial Average fund	0.27	0.44
Report owning investments with a concentration in industrial manufacturing or materials	0.08	0.27
Investment knowledge		
Mutual fund knowledge score (0-11)	4.32	3.06
Knowledge of sector return beliefs (0-3)	1.21	0.89
Prediction error for beliefs about S&P 500 growth (distance to true growth rate, in percentage points)	11.28	15.26
Device used to answer survey		
Computer	0.45	0.50
Tablet	0.12	0.33

[15] Due to some abnormal submissions by participants regarding their net wealth, submissions that were below the 1st percentile (-$200,000) of reported net worth or above the 99th percentile ($6,000,000) of reported net worth were set to -$200,000 and $6,000,000, respectively. These trimmed values were then used in all analyses instead of the original abnormal submitted values.

Variable	Mean	Std. dev.
Mobile	0.41	0.49
Other	0.01	0.12

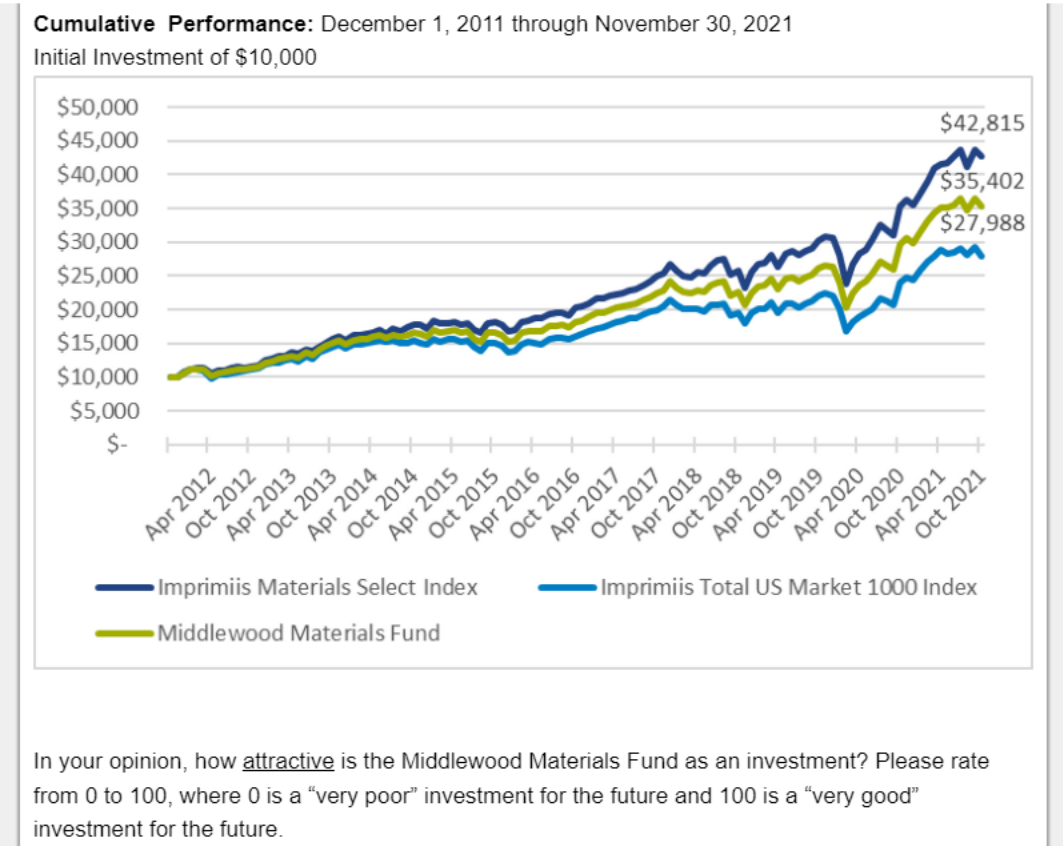
3.4 Experimental Design and Measures

Participants began the experiment by answering survey questions about their household's financial management, beliefs about the 10-year historical performance of the S&P 500 index (gone up, gone down, or stayed the same, with an annual percent change follow-up), and how specialized sectors (technology, healthcare, and manufacturing) had performed over that period. We used these variables to assess general familiarity and knowledge of investing.

Next participants considered a scenario in which they had inherited $15,000 in the Middlewood Materials Fund, a fund that "invests in companies that collect and process raw materials" and charges expenses of 0.25% per year.[16] They were randomized to one of eight conditions that varied in terms of graphical presentations (see Figure 4 for an example; full stimuli in Appendix C). The "fund only" condition included a cumulative performance graph that displayed only the Middlewood Materials Fund, and acted as the control condition for the majority of the analyses. Six of the other conditions included additional benchmarks (narrow only, broad-based only, or both), that varied in placement (if the respective benchmark(s) were outperforming or underperforming the Middlewood Fund). The final "no graph" condition did not include any performance information, and was designed to measure how historical performance information influenced beliefs about the Middlewood Fund's future performance. In the two conditions where both benchmarks were shown, participants were also randomly assigned to see narrative text, or not, to assess the relative impact of explaining the benchmarks' content.

[16] This expense ratio is at the 13th percentile for funds with that specialization since 2000 (which are observed every year for each fund that exists in that year), based on the Morningstar data.

Figure 4. Example of experimental stimuli for a condition with both benchmarks.



Participants reported their subjective evaluations of the fund's attractiveness ("In your opinion, how attractive is the Middlewood Materials Fund as an investment?" 0 = Very poor to 100 = Very good) and were asked to explain their ratings in a few sentences (open-ended text box). They were also asked for their evaluation of the fund's historical performance ("In your opinion, how well do you think the Middlewood Materials Fund performed over the past 10 years?" 0 = Very poor performance to 100 = Very good performance), their subjective assessment of the fund's riskiness ("In your opinion, how risky is the Middlewood Materials Fund as an investment?" 0 = Not at all risky to 100 = Extremely risky) and information confidence ("If you were making an investment decision today, how confident are you that you have enough information to make decisions about the Middlewood Materials Fund?" 0 = Not at all confident to 100 = Extremely confident).

To provide a behavioral measure of investment activity, participants were next asked to make three allocation decisions, in which they could allocate a $15,000 investment between the Middlewood Materials Fund or an account with a guaranteed return (with interest rates of 6%, 4%, and 2%). All participants were informed that subset of participants would be paid based on how much money they had remaining after a 6-month period; for instance, if they ended up with $12,000 remaining, they could receive a bonus payment of $120. These allocation decisions

were followed by four survey questions that could provide insight on why participants chose to invest (or not), such as "I am not interested in investing in a materials fund."

Consistent with experiments testing informational interventions on subjective expectations (e.g., Armantier, Nelson, Topa, van der Klaauw and Zafar, 2016; Armona, Fuster, and Zafar, 2016), participants were asked to assign probabilities of various Middlewood Materials Fund price movements over the next 6 months. Specifically, they were asked to assign a percent chance to each of 6 price bins for a $100 investment in the fund (ranging from being worth "$90 or less" to being worth "$130 or more").

After asking for beliefs about the fund's performance, we asked a series of multiple choice questions about participants' interpretation of the graphs (e.g., "In terms of total returns from December 2011 to November 2021, how did the Middlewood Materials Fund perform relative to the materials sector?"), their preferences for benchmark information, and other beliefs about the graphs that were shown (e.g., "This graph was designed to make the Middlewood Materials Fund look good"). Many of the statements about the graphs were drawn from the qualitative interviews conducted with participants prior to the study, as described above. We chose to administer these statements to assess the frequency of lay beliefs about investment performance graphs. Finally, the experiment concluded with background questions about participants, including their mutual fund knowledge, risk preferences, and wealth. We collected device type to control for the possibility that respondents using mobile devices could not see the graphs.

4. Predictions and Decisions

Our two primary outcome measures may lead to two different interpretations by study participants because they differ in terms of reference settings. To evaluate the attractiveness of the Middlewood Materials Fund, participants could draw on outside knowledge or the stimulus presented. In such circumstances, participants may ignore the graphical information presented entirely, or they may evaluate the fund against the graphical information; for example, they could compare the fund's performance against the benchmarks shown. Standard economic theory does not provide much guidance on how benchmarks should affect subjective evaluations: in a strict rational-expectations formulation, well-informed rational agents would find benchmarks ignorable because they would have imbibed sufficient outside knowledge prior to the experiment to form a basis for an evaluation. Thus, there should be no difference across conditions in subjective evaluations. Yet, weaker versions of a standard model could introduce a role for benchmarks such as through search costs or Bayesian updating.[17]

The participation outcome provides a slightly narrower scenario to evaluate. In making an allocation decision, participants should evaluate the fund against the guaranteed rate of return based on expectations of the fund's future performance and their risk preferences. However, in

[17] Much research suggests that in this decision-making domain, many investors may lack knowledge consistent with the strictest models one could consider (see, for example Scholl and Fontes, 2021).

the strictest rational model, the graphical stimuli should not play a role because the benchmarks do not affect the choice that the participant is making. That is, the participant is presented only with the choice of the risky gamble between the guaranteed rate of return and the uncertain outcome of the fund. They do not have the opportunity to liquidate the investment amount to pursue an outside option, and the timing of the payment will come six months after their decisions regardless of their investment. For the graphical information to matter in a standard economic model, participants would need to update their expectations of the fund's performance based on the graphical presentation, or change their risk sensitivity. For example, the graphical presentation could change a participant's expectations of the fund's future volatility.

5. Empirical Results

We estimate the effects of information by regressing our outcomes of interest on variables representing the information presented. The primary estimating equation for subjective attractiveness ratings is:

$$(1)\ (Attractiveness)_i = \beta_0 + \beta_1(benchmark\ below\ fund\ only)_i + \beta_2(benchmark\ above\ fund\ only)_i + \beta_3(two\ benchmarks)_i + \beta_4(no\ graph)_i + \beta_5(narrative)_i + \varepsilon_i$$

While estimation of the allocation to the fund when given an option to invest in a risk-free asset with guaranteed return a is provided by:

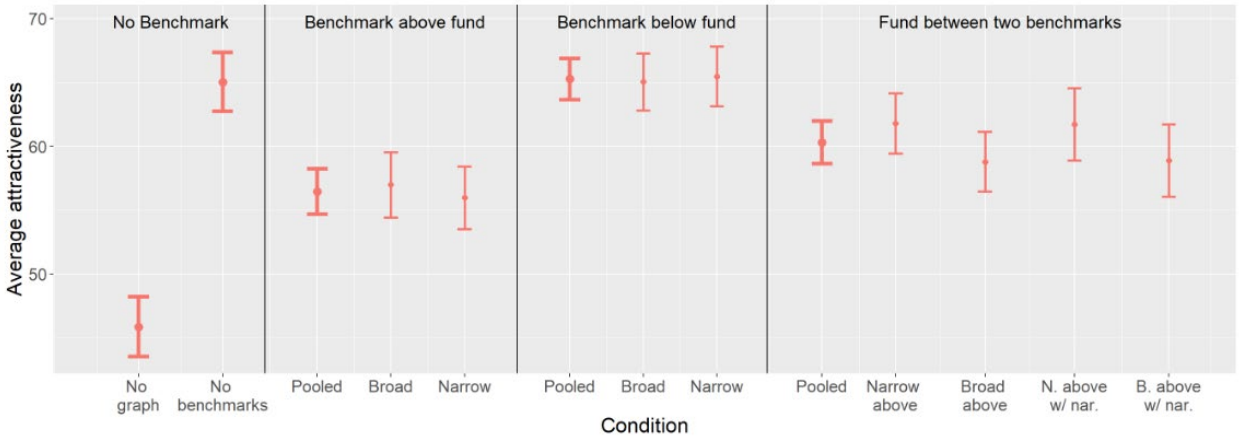
$$(2)\ (Allocation)_{i,a} = \beta_0 + \beta_1(benchmark\ below\ fund\ only)_i + \beta_2(benchmark\ above\ fund\ only)_i + \beta_3(two\ benchmarks)_i + \beta_4(no\ graph)_i + \beta_5(narrative)_i + \sum_a \delta_a(guranteed\ return = guranteed\ return_a)_{i,a} + \varepsilon_{i,a}$$

Equation (1) is estimated using robust standard errors, whereas in equation (2), responses are clustered by respondent, as each participant provides responses at guaranteed returns of 2, 4, and 6 percent.

We first study effects on attractiveness evaluations in the online experiment. This variable is useful because it shows whether the interventions had any impact on study participants' overall impressions of the fund. We then estimate any effects in an incentive-compatible choice, which allows us to determine if the reported subjective evaluations seep into behavioral differences across stimulus conditions.

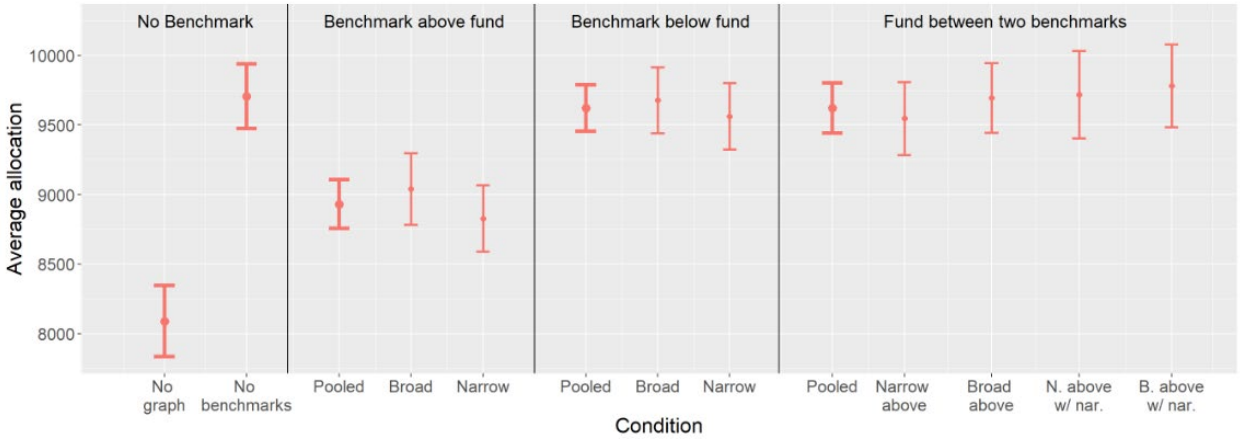
Figures 5 and 6 display point estimates and confidence intervals of participants' subjective attractiveness ratings and allocation decisions across conditions. These plots provide considerable intuition corresponding to formal estimates presented in Appendix E. Our research design varies two main features vis-à-vis the base (no benchmark) condition: the relative position and number of benchmarks, and the selection and positioning of either the broad or narrow benchmark. The relative position focuses our attention on participants' reaction to benchmark performance relative to the fund. Broad vs. narrow variation enables us to examine if participants distinguish between broad and narrow benchmarks – for example, if participants respond differently to a broad benchmark outperforming the fund rather than a narrow benchmark outperforming the fund. The plots group the stimuli in terms of relative position of the benchmarks in the respective stimulus (specifically, graphs with no benchmarks, a benchmark above the fund, a benchmark below the fund, and the fund between two benchmarks). Where there was variation by benchmark type, we provide estimates based on solely the relative position ("pooled"), and separate out estimates for the broad and the narrow benchmark. This portrayal highlights the fact that participants' responses tend to be driven more by relative position than whether or not the benchmark presented is broad or narrow.

Figure 5. Average attractiveness evaluation by condition.



Note. This figure provides point estimates and confidence intervals of participants' attractiveness ratings across conditions. The five thicker confidence intervals represent averages for the following conditions: no graph, no benchmark, single benchmark above fund, single benchmark below fund, and two benchmarks. In contrast, the eight thinner confidence intervals represent mean values that distinguish between broad and narrow benchmarks, as well as narrative text ("w/ nar.").

Figure 6. Average allocation to the Middlewood Fund (vs. fixed return).



Note. This figure provides point estimates and confidence intervals of participants' allocations to the Middlewood Materials Fund across conditions. The five thicker confidence intervals represent averages for the following conditions: no graph, no benchmark, single benchmark above fund, single benchmark below fund, and two benchmarks. In contrast, the eight thinner confidence intervals represent mean values that distinguish between broad and narrow benchmarks, as well as narrative text ("w/ nar.").

Figure 5 provides confidence intervals across conditions for the subjective attractiveness outcome. The figure provides evidence that benchmark presentation affected participants' subjective evaluations of the fund. When participants received a graph depicting a single benchmark that outperformed the fund, they provided ratings of the fund that were approximately 8.5 points lower than participants in the excluded condition (performance graph with "No Benchmarks"). This difference is significant at the 99.9% level. The two benchmark conditions affected subjective evaluations in a more muted way: participants' evaluations were 4.8 points lower than the base condition, perhaps suggesting that participants were affected both by the reference value that outperformed and the reference index that underperformed the fund. The two-benchmark p-value is 0.013.

In contrast to the benchmark that outperforms the Middlewood Fund, a single benchmark that performs *worse* than the fund did not affect participants' subjective ratings vis-à-vis the base condition ($p = 0.878$). Results suggest that participants' evaluations were not affected by this graphical presentation.

Participants in the no graph condition had substantially lower subjective appraisals of the Middlewood fund. The average participant in this condition rated the fund 19.2 points lower than in the base condition ($p < 0.001$).

3.2 Effects on Incentivized Behavior: Allocation to Middlewood

Figure 6 illustrates that allocations were highest overall in the no benchmarks condition and that allocations varied somewhat with the relative position of the benchmark. The figure also highlights that participation decisions were influenced by the graphical presentation of certain conditions, suggesting that benchmark presentation is consequential for investment decisions.

In particular, participants assigned to the single benchmark above conditions ("Benchmark Above Fund") exhibited significantly different participation behavior than participants viewing benchmarks in other positions. Participants allocated an average of $779 less to the fund in this condition than they did in the no benchmark condition ($p < 0.001$). In other words, after controlling for the guaranteed return offered, participants viewing a single benchmark line outperforming the fund were more likely to allocate money into the guaranteed return rather than invest in the fund.

In contrast, behavior in response to the pooled single benchmark below fund conditions ("Benchmark Below Fund") was indistinguishable from the no benchmarks condition. While the pooled point estimate directionally indicates slightly lower participation of about $87, we cannot reject the null of no difference with the base condition. In other words, this difference is statistically indistinguishable from the base condition and attributable to chance variation. Results for the "Fund between Two Benchmarks" conditions are also similar to the no benchmark condition. Although the statistically insignificant point estimate indicates directionally lower allocations than the base condition, there is a muted response relative to conditions where participants view a single benchmark outperforming the fund.

Finally, the no graph condition leads to substantially lower allocation in the fund than any of the graphical presentation conditions. Participants in this condition allocated $1,618 less to the fund ($p < 0.001$).

3.3 Broad vs. Narrow Benchmarks

Figures 4 and 5 (and Appendix Table E) also break out the relative performance conditions of the benchmark (the "pooled" results) in a way that allows us to make distinctions between broad and narrow benchmarks. That is, these figures allow us to assess whether participants react differently when viewing a broad benchmark versus a narrow benchmark, and to what extent attractiveness ratings and allocations might be influenced differently by the relative position of each. While the point estimates do differ slightly within each of the positional conditions in the figures, these differences are not statistically significant, and there is little evidence that participants responded differently to the hypothetical broad and narrow Imprimiis benchmarks we constructed. Figures 4 and 5 emphasize that – at least in the experimental set-up we have explored – the relative position of the benchmark versus the fund appears to be much more consequential than whether the benchmark presented is narrow or broad.

3.4 Subgroup Analysis

One potential mechanism for an effect of benchmarks on participants could be that high sophistication participants might ignore the benchmarks because of outside knowledge of the marketplace; if high sophistication participants roughly know the historical returns of stocks or a given sector, they might essentially impose their own reference value – leading to no effect of benchmarks in different positions. By contrast, less sophisticated investors might use the benchmarks to provide context that would enable them to make an assessment of the attractiveness and desirability of the fund given their lower level of familiarity with the investment space.

Alternatively, it could be the case that sophisticated investors use benchmarks as an indicator of relative performance because they understand that benchmarks provide context. In contrast, less sophisticated investors might find benchmarks confusing and choose to ignore them. In this case, we would expect a greater response to benchmarks among sophisticated (vs. unsophisticated) investors.

3.4.1 Investor Subgroup Variable Creation

To explore whether effects on attractiveness evaluations and allocations vary by participant characteristics, we next classify each participant into one of the following three categories: non-investors, retirement-only investors, or independent investors. We constructed these subgroups with the expectation that independent investors would have the highest levels of investment sophistication and experience with regard to funds among a retail (as opposed to institutional) investor population. Additionally, we expected that non-investors would have lower levels of investment knowledge and experience than retirement-only investors. This classification, and these expectations, are based on prior research that distinguishes investors' sophistication (e.g., Chin, Scholl, and VanEpps, 2021; Scholl and Fontes, 2021).

To determine a participant's sophistication level, we used four pre-experiment screening questions from Chin, Scholl, and VanEpps (2021). Participants who had an employer-sponsored retirement plan but no ability to choose among investments in the plan (as in the case of most pensions), as well as participants who reported no investments, were classified as non-investors. Anyone who chooses investments in their employer-sponsored retirement plan or has retirement accounts outside of an employer-sponsored plan (e.g., an individual retirement account), but no other stock, bond, mutual fund or ETF investments outside of a retirement account, was classified as a "retirement-only" investor. Finally, anyone who reported having investments in stocks, bonds, mutual funds, or other securities outside of a retirement account (e.g. in a brokerage account, or in actual stock certificates) was considered an "independent" investor. This last group likely includes respondents with a retirement account as well. Ultimately, these classifications are imperfect proxies of investment experience, but we believe this classification helps to contextualize participants' level of investment experience and fund knowledge.

To provide additional context on the three subgroups, Table 3 shows the breakdown of four variables we might expect to correlate with investor sophistication: (1) the deviation between beliefs about historical stock market performance (measured by the S&P 500 index) and actual performance; (2) beliefs about how various sectors performed relative to the overall US stock market (better/worse/about the same/I don't know) – this variable is the count of the number of sectors with relative performances that the respondent answered correctly out of 3 sectors; (3) mutual fund literacy, as assessed by a validated scale developed in prior research (Scholl and Fontes, 2022); and (4) responses to Ipsos' profile questions about whether the respondent owns mutual funds or ETFs. As shown in the table, all four of these variables provide a consistent pattern between subgroups. Non-investors are the least sophisticated, as they have the most inaccurate beliefs, lowest mutual fund literacy, and lowest levels of fund ownership. Note that about 5% of our "Non-Investors" report owning mutual funds or ETFs – the difference here largely reflects data previously collected by Ipsos and questions we asked directly in our survey. "Independent" investors have the highest levels of sophistication, and retirement-only investors fall in between. We now proceed to examine responses to the experiment by these subgroups.

Table 3. Measures of investor sophistication by subgroup.

	Full Sample	**Independent Investor**	**Retirement Only Investor**	**Non-Investor**
Difference between true stock market return and return belief (ppt.)	11.3 (15.3)	8.3 (13.5)	10.3 (13.2)	16.5 (17.7)
Sector performance score (0-3)	1.21 (0.89)	1.41 (0.84)	1.19 (0.89)	0.94 (0.89)
Mutual fund literacy score (0-11)	4.32 (3.06)	5.68 (2.85)	4.10 (2.82)	2.56 (2.53)
Whether respondent owns mutual fund(s) or ETF(s)	0.417 (0.493)	0.698 (0.459)	0.363 (0.481)	0.0573 (0.232)

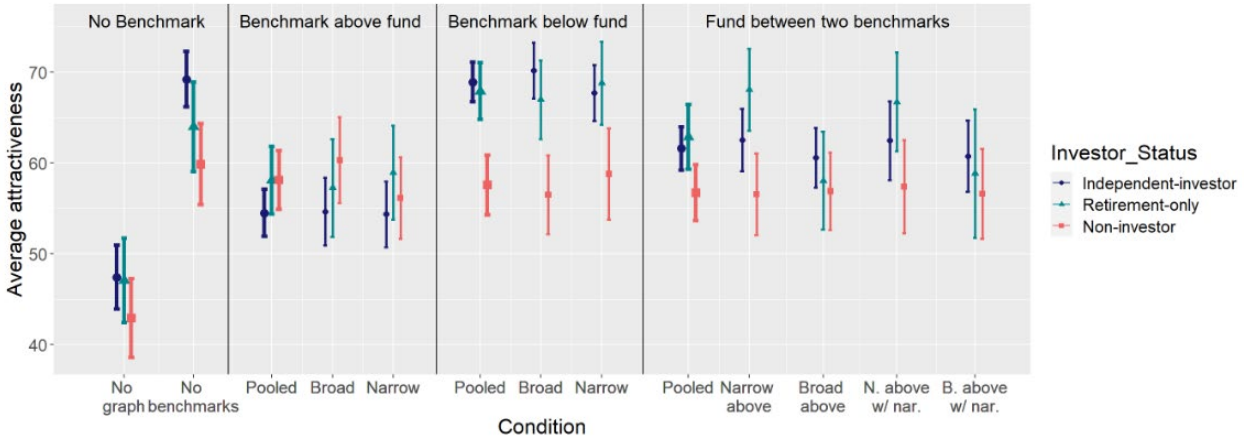
Note. Means and (standard deviations) shown.

3.4.2 Attractiveness Evaluations by Investor Subgroup

Figure 7 and Table 4 provide subgroup estimates based on investor status. They paint a surprisingly different story than the proposition that more sophisticated individuals are less responsive to benchmarks. Instead, there is evidence that the most sophisticated are the most responsive. Non-investors (Column 3 of Table 4, and Figure 6) demonstrate no differential reaction in Attractiveness score to any of the positional conditions – in the graphical presentation, this group clearly does not assign differential ratings based on their condition. Retirement-only investors (Column 2), exhibit a small, marginally significant decrease in

Attractiveness score of about 5.9 points to having a benchmark above the fund. Yet, independent investors (Column 1) respond differentially based on performance presentation. Independent investors' ratings of fund attractiveness decrease by 14.7 and 7.7 points in the "Benchmark above fund" conditions and the two benchmarks conditions, respectively. As per Figure 6, there is little evidence that participants varied systematically in their evaluations for broad and narrow benchmarks although retirement-only investors differentiated within the two benchmark condition based on whether the narrow benchmark was above or below the fund.

Figure 7. Attractiveness evaluations by investor subgroups.



Note. Figure presents group means for each investor subgroup and corresponding 95% confidence intervals.

Table 4. Fund attractiveness by investor subgroup.

	Baseline	Independent Investors	Retirement-Only Investors	Non-Investors
	(1)	(2)	(3)	(4)
No graph, no benchmarks	-19.167***	-21.770***	-16.911***	-16.934***
	(1.667)	(2.350)	(3.415)	(3.155)
Single benchmark above fund	-8.574***	-14.70***	-5.892*	-1.741
	(1.474)	(2.030)	(3.122)	(2.798)
Single benchmark below fund	0.220	-0.311	3.904	-2.293
	(1.426)	(1.903)	(2.952)	(2.809)
Two benchmarks	-4.781**	-7.674***	-1.014	-3.823
	(1.921)	(2.584)	(3.941)	(3.869)
Any narrative	0.045	0.0976	-0.167	0.952
	(1.838)	(2.547)	(3.795)	(3.613)
Constant	65.036***	69.195***	63.985***	59.857***
	(1.165)	(1.547)	(2.487)	(2.262)
Observations	4,226	1,906	998	1,322
R^2	0.047	0.077	0.052	0.029

Adjusted R^2	0.046	0.075	0.047	0.026

Note. Robust standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01

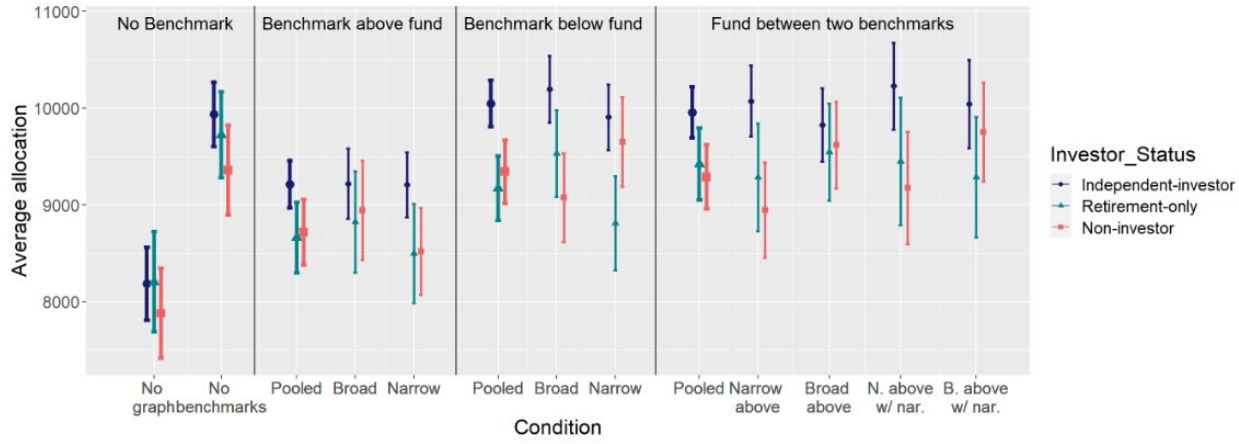
3.4.3 Allocation Decisions by Investor Subgroup

In terms of allocation behavior, Figure 8 and Table 5 suggest a similar pattern to those on evaluations, with some subtle differences. Table 5 Column (4) suggests that the only effect of presentation differences for non-investors came from the comparison between the reference case (a graph with fund performance and no benchmarks), and the No Graph condition. All other presentations did not differ significantly from the no benchmark condition.

In contrast, the purportedly more sophisticated participants, retirement-only and independent investors, allocate substantially less to the Middlewood Fund in the single benchmark above condition. For retirement-only investors, this difference amounts to $1,063 on average across the three guaranteed returns, while for the independent investors it amounts to $725 less. Curiously, some of the directional values of the non-statistically significant coefficients are somewhat at odds with results reported above; for example, retirement-only investors and non-investors allocated less on average in the single benchmark below condition. As with evaluations, there is little evidence in Figure 8 that participants differentially responded to broad or narrow conditions – rather, the primary driver appears to have been the relative position of the benchmark line.

Overall, the implication is that more sophisticated participants have higher responsiveness to differential benchmark presentations than their less sophisticated peers who may not have sufficient context or understanding to make use of the benchmarks.

Figure 8. Allocation decisions by subgroup.



Note. Figure presents group means for each investor subgroup and corresponding 95% confidence intervals.

Table 5. Allocation decisions by investor subgroup.

	Baseline	Independent Investors	Retirement-Only Investors	Non-Investors
	(1)	(2)	(3)	(4)
No graph, no benchmarks	-1,618.666***	-1,753.034***	-1,523.308***	-1,481.129***
	(275.457)	(385.088)	(542.812)	(543.131)
Single benchmark above fund	-779.103***	-725.064**	-1,063.698**	-646.958
	(230.886)	(316.964)	(457.308)	(466.586)
Single benchmark below fund	-87.487	108.297	-555.092	-16.575
	(226.954)	(313.473)	(438.587)	(461.706)
Two benchmarks	-375.079	-371.769	-208.001	-552.441
	(322.007)	(445.992)	(609.480)	(666.342)
Any narrative	414.313	573.179	-149.772	661.922
	(313.927)	(436.409)	(615.946)	(633.289)
Guaranteed Return of 4%	-1,341.813***	-1,686.459***	-1,303.103***	-871.129***
	(44.847)	(66.958)	(89.822)	(79.773)
Guaranteed Return of 6%	-2,512.655***	-3,273.571***	-2,468.919***	-1,438.529***
	(63.700)	(97.854)	(125.968)	(105.040)
Constant	10,992.060***	11,588.230***	10,981.330***	10,129.840***
	(186.073)	(255.915)	(357.443)	(379.951)
Observations	12,434	5,629	2,944	3,861
R^2	0.054	0.093	0.052	0.022
Adjusted R^2	0.053	0.090	0.046	0.017

Note. Robust standard errors clustered by participant in parenthesis. *p<0.1; **p<0.05; ***p<0.01

3.5 Deviations from expected utility maximizing allocations

We now explore by how much allocations differ from the allocations that would maximize expected utility given the participants' beliefs about the Middlewood fund's future returns. We assume that participants have utility functions that exhibit constant relative risk

aversion (CRRA) and infer their coefficient of risk aversion from a survey question.[18] We then find each participant's expected utility maximizing allocation by numerically integrating expected utility given their coefficient of risk aversion and the fitted distribution for their beliefs (as described above) for each possible allocation from zero to $15,000 in one dollar increments. The utility maximizing allocation is the value that maximizes this grid search. Since the survey question provides us with a range for the coefficient of relative risk aversion, we calculate a range of utility maximizing allocations. We record the deviation from the utility maximizing allocation as zero if the allocation falls in the range and as the minimum distance to the range if it falls outside of it.

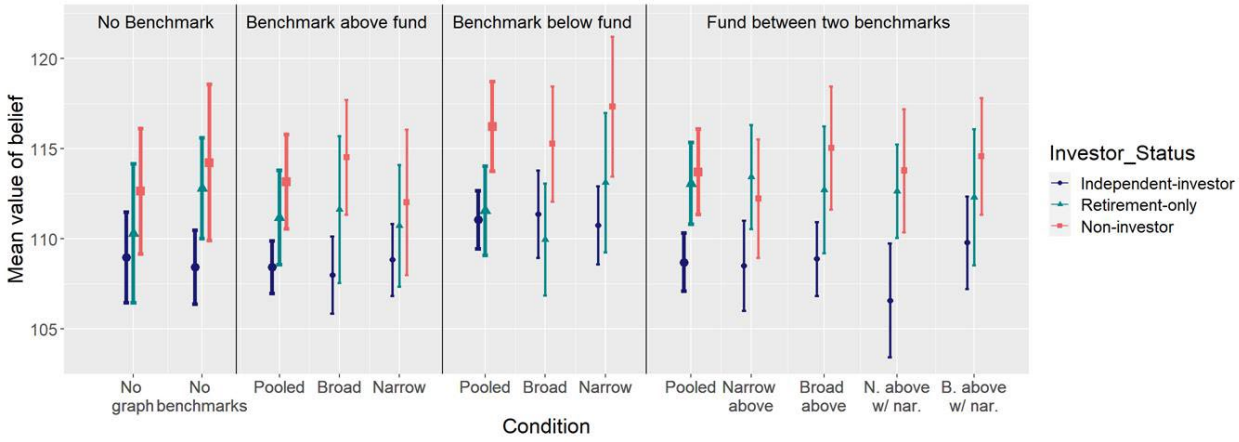
With this approach, the distance from the utility maximizing allocation is thus driven by the participant's allocation to the fund and the participant's beliefs about the fund's future returns. To understand changes in beliefs, we first present Figure 8, which displays the mean of the fitted distributions for beliefs by investor type.[19] Overall, Figure 8 demonstrates that non-investors have less responsive beliefs and exhibit more variation in their responses. However, the more surprising overall pattern is that expectations – meaning participants' projections of actual future performance of the fund – are largely unaffected by condition. Statistical tests only find a significant difference between the single above and single below conditions (diff=-2.19, p-value=0.01). We thus find limited evidence that expectations are updating, but strong evidence that allocation varies by condition.[20]

[18] In the survey, we elicit CRRA by asking participants to choose between one of eight lotteries, varying from the least risky, which pays $1.60 for sure, to most risky, which offers a 50% chance at $4.40. The inferred coefficients of relative risk aversion, denoted as r, for these lotteries are: for lottery 1, $3.9437 < r$; for lottery 2, $1.3199 < r < 3.9437$; for lottery 3, $0.8052 < r < 1.3199$; for lottery 4, $0.5748 < r < 0.8052$; for lottery 5, $0.4375 < r < 0.5748$; for lottery 6, $0.3404 < r < 0.4375$; for lottery 7, $0 < r < 0.3404$; for lottery 8, $r < 0$.

[19] Specifically, the survey captured beliefs about the Middlewood Materials Fund's return over the next six months by asking participants to assign probabilities across six bins corresponding to different ranges of returns. Using these assigned probabilities over bins, we fit probability distributions to model their beliefs following Engelberg, Manksi, and William (2009). Engelberg, Manksi, and William fit the parameters of a unimodal distribution, either generalized beta or isosceles triangle depending on the number of bins that were covered, to match the probabilities that were reported in bins. A small percentage of participants (two percent) reported beliefs that cover non-adjacent bins, for example a .5 probability that returns are between 10 and 20% and a .5 probability that they are 30% or more. Engelberg, Manksi, and William do not discuss the probability distribution for this situation of non-adjacent bins, as no participants in their sample report beliefs like this. We use a piece-wise uniform distribution, which evenly distributes the probability over the interval to which it was assigned.

[20] One possibility is that our expectations elicitation bin sizes are too wide to pick up expectations movements.

Figure 8. Expected return beliefs by investor type subgroups.



Note. Figure presents group means for each investor subgroup and corresponding 95% confidence intervals.

Moving on to modeled utility maximizing allocations, the first column of Table 6 presents results from regressing the deviations from the utility maximizing allocation on the benchmark presentation (single above, single below, two benchmarks, and no graph). The smallest deviation from utility maximization is found when the performance graph is presented with no benchmarks. The greatest deviation is for the condition that did not see a graph ($1,068 less than the condition that saw a graph without benchmarks), which had the lowest allocations but beliefs that were similar to the condition without benchmarks. The deviations for presentations with a benchmark under- and over-performing the fund are similar ($759 and $831 less than the condition without benchmarks, respectively).

When separating the participants by investor status in columns (2) through (4) of Table 6, we see that non-investors exhibited a general reluctance to invest, allocating too little to the fund given their beliefs, but this under-investment does not vary much by condition. This is unsurprising given that this group allocated far less to the Middlewood Materials Fund in every condition. Independent investors had allocations that more closely matched their utility maximizing allocations. Even though beliefs for independent investors varied more by condition, their allocations to the fund are largely consistent with these beliefs (with the exception of allocations in the condition that did not see a performance graph).

Table 6. Regression Results for Distance from Optimal Allocations.

	Baseline	Independent Investors	Retirement-Only Investors	Non-Investors
	(1)	(2)	(3)	(4)
No graph, no benchmarks	-1,067.980**	-1,224.721**	-668.059	-1,022.091
	(418.894)	(593.795)	(869.752)	(793.313)

Single benchmark above fund	-831.184**	-717.906	-508.820	-1,329.345**
	(346.666)	(471.920)	(759.599)	(661.779)
Single benchmark below fund	-759.209**	-795.477*	-265.554	-1,131.076*
	(345.584)	(472.269)	(752.373)	(660.529)
Two benchmarks	-257.692	-344.906	405.507	-757.452
	(484.292)	(675.251)	(1,040.035)	(912.423)
Any narrative	-381.058	14.505	-952.807	-363.975
	(468.577)	(670.241)	(983.290)	(862.567)
Constant	-61.357	574.511	-669.159	-597.996
	(279.064)	(379.069)	(618.915)	(529.677)
Observations	11,490	5,376	2,783	3,331
R^2	0.002	0.003	0.002	0.004
Adjusted R^2	0.002	0.001	0.000	0.002

Note. Standard errors clustered by participant in parenthesis. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

3.6 Search Effort

As discussed in the introduction, researchers examining mutual fund choice often discuss the search costs associated with finding a fund. After viewing the benchmark presentation and providing subjective ratings (prior to allocation and expectations elicitation tasks), respondents were asked about potential search behavior. The elicitation question was: "Let's say you had 60 minutes to spare. How much of it would you spend researching the Middlewood Materials Fund, searching for other funds, or doing something else (like watching TV)?" Respondents allocated 60 minutes to search for more information about the Middlewood fund, search for other funds, or doing something else – to reduce error, "doing something else" was automatically computed as the residual of the other two values.

One pathway through which benchmarks could affect evaluations and allocation decisions is through search costs. With an appropriately chosen benchmark that approximates true performance in the sector (e.g. a sector average or factor model) it is conceivable that investors could reference the benchmark as a means of evaluating past performance of the fund. Under this framework, both low and high sophistication investors could use a benchmark as a shorthand to avoid costly search activities. They would also not have to construct their own reference or comparison points.

Table 7 presents results of self-declared search effort. The only condition that differs from the baseline no benchmark condition is the single benchmark above condition. In this condition, participants reduced search on the Middlewood Materials Fund by about 2.6 minutes, and increased search for other options by about 1.7 minutes. All other conditions yielded no

differences in the level of search, relative to the condition where the fund performance, but no benchmarks were shown.

Table 7. Regression Results for Self-Reported Search Effort by Experimental Condition.

	Middlewood Info	Other Options
	(1)	(2)
No graph, no benchmarks	-0.828	-0.434
	(1.048)	(0.811)
Single benchmark above fund	-2.582***	1.666**
	(0.842)	(0.688)
Single benchmark below fund	-1.459*	0.713
	(0.847)	(0.672)
Two benchmarks	-0.939	1.196
	(1.147)	(0.925)
Any narrative	-0.406	0.119
	(1.088)	(0.905)
Constant	21.389***	14.260***
	(0.701)	(0.545)
Observations	4,196	4,196
R^2	0.003	0.003
Adjusted R^2	0.002	0.002

Note. Robust standard errors shown in parentheses. *p<0.1; **p<0.05; ***p<0.01

On one hand, the changes in search for the "benchmark above" condition are somewhat consistent with a search cost role for benchmarks in that there is some difference in search behavior. Yet, this pattern is not fully consistent with that view. A single benchmark outperforming the fund appears to signal to participants that they could find better investment performance elsewhere. However, the need for increased search effort dissipates in the two benchmark case, even though the graph continues to display investment options with superior historical performance. In this condition, the lack of change in search behavior suggests that having intermediate performance may be "good enough" for participants. Possibly, the presence of a benchmark below (either in the single benchmark case or in the two benchmark sandwich case we pursue) is sufficient to forestall motivation to search.

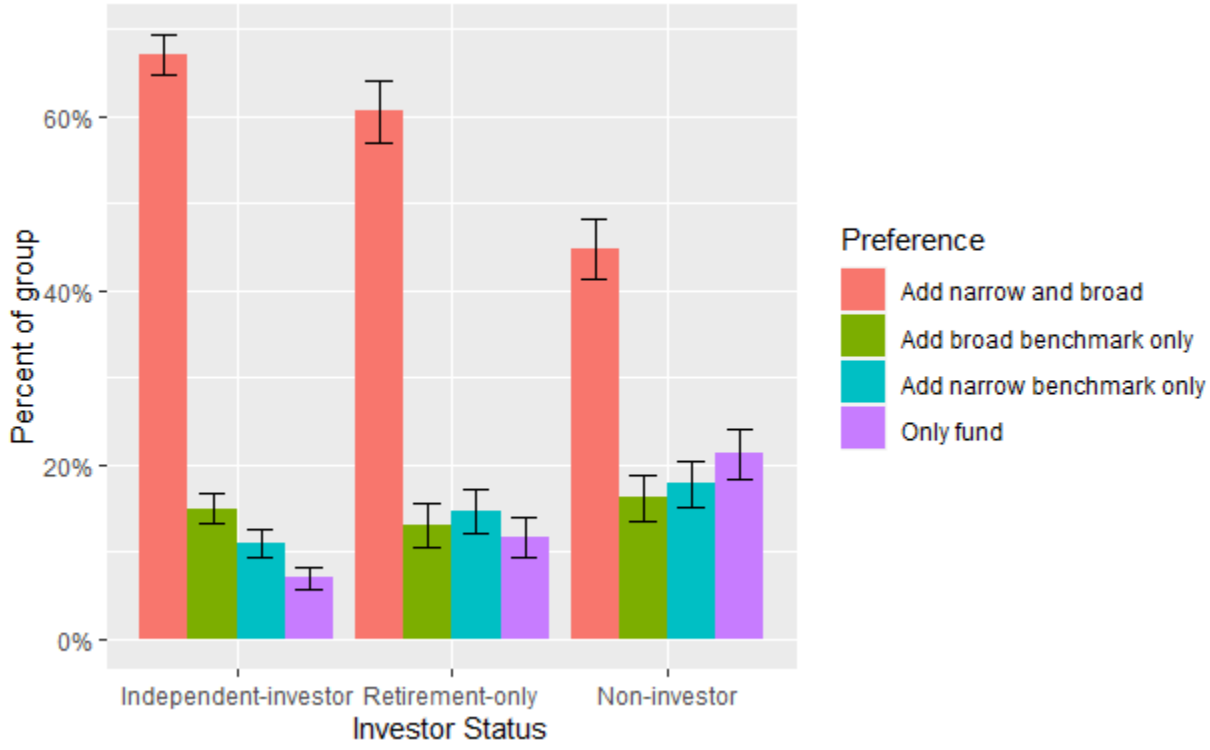
The issue of the intention to search seems especially important in the context of funds' discretion over benchmarks. A carefully selected benchmark might encourage investors to stop looking further.[21] Moreover, many investors may be effectively defaulted into a shortlist of funds based on retirement plan menus and other menus. With such a shortlist, a fund outperforming its benchmark may be sufficient information for the investor to select it, not knowing, or not fully factoring in, that the benchmark itself may have been selected for that particular effect. We further explore benchmark choice in Section 7 below.

[21] Scholl, Silverman and Enriquez (2021) develop such a model.

6. Survey Responses by Investor Subgroup

Our survey instrument collected additional information intended to produce nationally representative survey responses. Figure 9 provides survey responses to a question about participant preferences for a graph with one or two benchmarks. This data was collected by asking questions that allowed us to credibly determine preferences (beginning with a single preference question; additional follow-ups allowed participants to express a desire for additional information). Regardless of investor status, participants overwhelmingly expressed a preference for a graph with both narrow and broad benchmarks. This result can add context to the evidence arising from the experimental analysis described above.

Figure 9. Preferences for benchmarks.



Note. Bars denote +/- 1 standard error.

As mentioned in the Introduction (Section 1), there are different ways in which benchmarks may, or may not, affect investor decision-making. Given differences in behavior between investors of different levels of sophistication, in Figure 10 we also present agreement with different survey items answered by our participants. Specifically, the figure shows average agreement to each of nine statements for participants in each of the investor subgroups.[22]

[22] We expected and found no differences in beliefs by experimental treatment.

Conceptually, these statements are grouped into four categories, as reflected in the subpanels of the figure.

First, we assessed mistaken beliefs about benchmarks, as derived from our qualitative pilot. We asked whether the Middlewood Materials Fund would always have performance between the two benchmarks, whether the indexes were competitors to the Middlewood Materials Fund, whether the lower benchmark in the graph is included in the Middlewood Materials Fund, and whether the hypothetical materials index provides the average of investments in that sector (top panel of Figure 10). As shown, participants did not strongly agree with these statements, as the midpoint of the response scale was 3. Where there are differences between groups, however, non-investors tended to agree more with the statements that participants with higher investment sophistication.

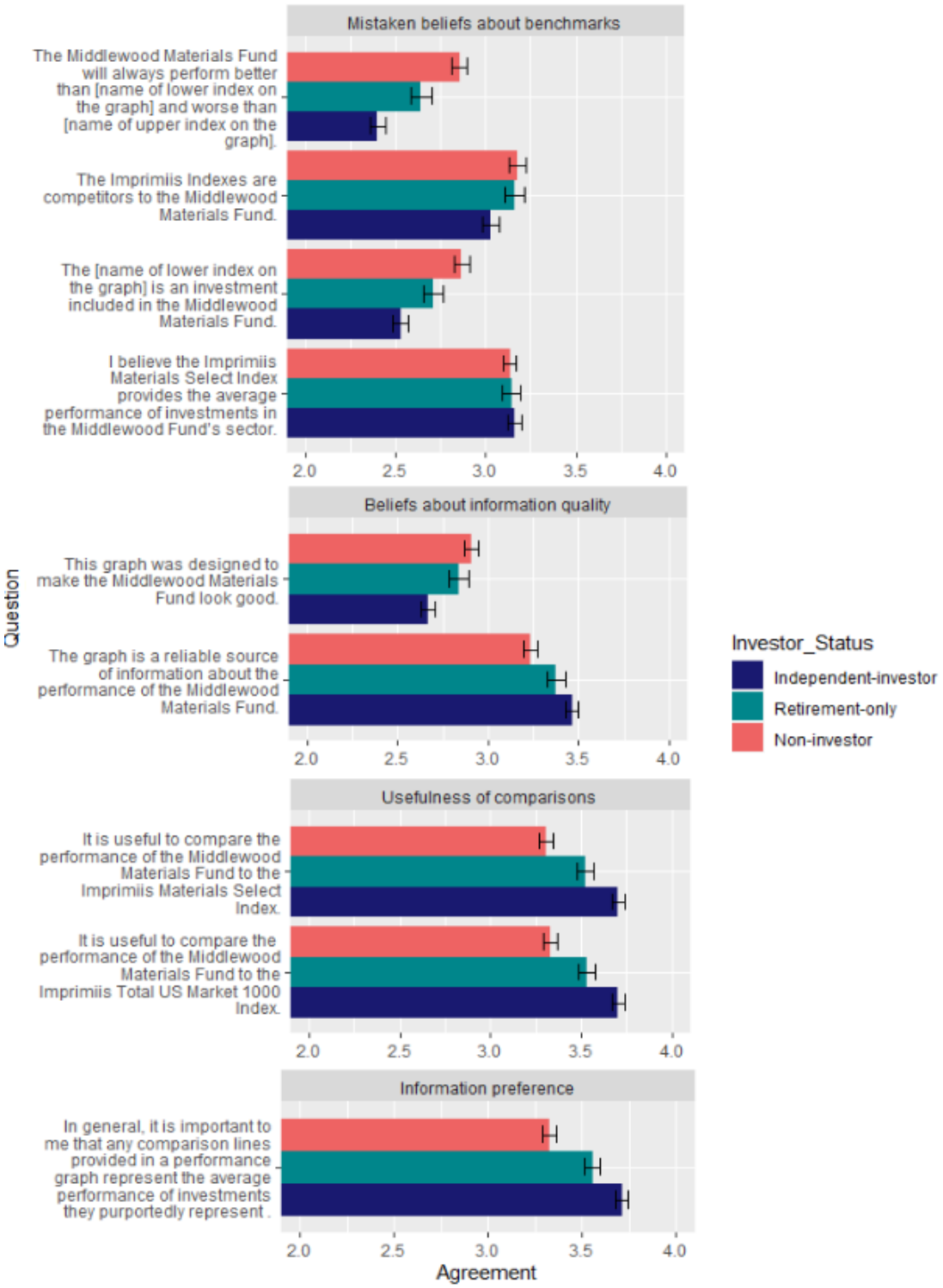
The second panel describes beliefs about the graphs shown in the study. Perhaps surprisingly, relatively sophisticated "independent" investors were the most likely to agree that the graph was a reliable source of information, and least likely to agree that it was designed to make the fund "look good." It is possible that these participants are more familiar with performance graphs and more willing to trust the information provided.

The third panel shows beliefs about how useful it is to compare the Middlewood Materials Fund to the hypothetical Imprimiis Indexes. As with the immediately preceding panel, there are differences by investor group; investors that are more sophisticated find the indexes more useful.

Finally, the last panel asks participants to state how important it is that the benchmarks represent average performance. Overall, participants agree with this statement as responses for all groups are above the midpoint of the scale. Additionally, more sophisticated investors tend to agree more.

Looking across the survey results, we conclude that the patterns of responses are largely consistent with the results of our behavioral experiment. In particular, non-investors, who are more confused by benchmarks (panel 1), trust the graphs less (panel 2), and find the benchmark comparisons less useful (panel 3), also respond less to benchmarks in the experiment than more sophisticated respondents (Section 5).

Figure 10. Survey items by investor subgroup.



Note. This figure shows averages by investor group. Bars denote +/- 1 standard error.

7. Analysis of Benchmark Performance Data

Thus far, we have described the frequency of benchmarks in mutual funds, and demonstrated experimentally that fund performance relative to a benchmark can affect evaluations and investment decisions. To complement these findings, we now return to a description of benchmarks that currently exist in the mutual fund industry. We ask: Does benchmark "fit" appear to vary with funds' choices of benchmarks?

Following a similar methodology to that used in Sensoy (2009) to identify benchmarks that match the fund's exposure to market-level fluctuations that are beyond the fund's control (i.e., market factors), for each fund we run regressions of funds returns on the returns of their benchmark. We use the average R-squared across the models as metric for benchmark fit.[23] ~~R-squared values range from 0 to 1, with 0 representing a case in which none of the variance in the dependent variable (in this case, the fund performance) is explained by the variance in the independent variable (in this case, benchmark performance).~~ The overall idea is that the benchmark with the most similar exposure to market factors as the fund will have returns that are the most highly correlated with the fund's.

Average R-squared values are displayed below in row 6 of Table 8 for primary and secondary benchmarks. In every sector, we find these values look similar between the primary and secondary benchmarks, with the average R-squared for primary benchmarks across all sectors similar to the average for secondary benchmarks (and in some sectors, the average R-squared is greater for primary benchmarks than secondary). This is surprising given that the justification for including secondary benchmarks is often to provide a more apples-to-apples comparison to the fund. On average, the primary benchmarks have exposure to factors that is more similar to the fund than the secondary benchmarks.

To further explore benchmark fit, we use Fama-French three-factor models. The traditional model seeks to capture performance based on three factors: the overall return on the market relative to the risk-free rate, the size of firms (SMB or "small minus big") and book-to-market values (HML or "high minus low"). Here, again following methodology in Sensory (2009), we examine the differences in performance between a fund and its benchmarks by implementing the following models separately for primary and secondary benchmarks for each fund:

$$R_{i,t} - R_{Bench,i,t} = \alpha_i + \beta_i(R_{M,t} - R_{f,t}) + s_i SMB_t + h_i HML_t + e_{i,t}$$

Where $R_{i,t}$ is fund i's return in month t and $R_{Bench,i,t}$ is the return of fund i's benchmark in month t. Our factor loadings quantify the extent to which performance differences between funds and their benchmarks can be explained by differences in exposure to these three factors. Similar to the logic used above, if a benchmark is a poor comparison, we would expect that deviations in

[23] R-squared values range from 0 to 1, with 0 representing a case in which none of the variance in the dependent variable (in this case, the fund performance) is explained by the variance in the independent variable (in this case, benchmark performance).

the factors would be more prevalent. As such, we calculate the frequency of statistically significant deviations in factor loadings across all primary and secondary benchmarks, and compare the rates for the two types. Funds with complete returns data for the fund and the funds respective benchmarks from beginning of 2017 to the end of 2019 are included.

In Table 8 rows 7 - 10 we show the percentage of funds which have statistically significant ($\alpha = 0.05$) differences in either direction in loadings between the fund and their benchmark for each of the three factors, split by sector and by primary and secondary benchmark. These results suggest that deviations between funds and their benchmarks in terms of Fama-French factors are common. With only a few exceptions, more than 50% of funds have differences in loadings across all three factors, when compared to both their primary and secondary benchmarks. Often this proportion is three quarters of funds or more. Additionally, in nearly every sector we find that the proportion of funds with significant differences in loadings with their secondary benchmarks is as high as or higher than the proportion for primary benchmark, when considering all three factors together.

In sum, little can be found in either of the two preceding lines of analysis to support a characterization of secondary benchmarks as being more informative about the fund's risk-adjusted returns (by better matching the fund's exposure to market factors) on average than primary benchmarks.

Table 8. Benchmark summaries by sector.

	Large Growth		Large Blend		Large Value		Divers. Emer. Mkts		Small Blend		Fn. Large Blend	
	Primary	Sec.	Primary	Sec.	Primary	Sec.	Primary	Sec.	Primary	Sec.	Primary	Sec.
% Funds w/ Benchmarks	99.68%	28.39%	98.95%	21.75%	98.53%	31.5%	100%	21.51%	100%	24.52%	100%	24.03%
# Unique Benchmarks	20	27	20	39	22	31	18	23	15	14	10	21
Benchmark Performance												
Avg. 3 year return	56%	58.15%	47.43%	41.43%	38%	41.03%	31.72%	32.72%	26.74%	21.99%	31.2%	29.53%
SD of 3 year return	16.22	13.67	9.25	13.82	10.23	9.72	6.01	7.45	12.91	4.93	4.03	5.75
Avg. volatility	3.56	3.66	3.47	3.37	3.41	3.41	3.77	3.88	4.43	4.6	3.2	3.19
Benchmark Fit												
Avg. R^2 of fund on benchmark	0.85	0.89	0.9	0.89	0.86	0.9	0.85	0.84	0.85	0.87	0.91	0.91
% funds diff. loading on market	76%	72%	67%	82%	62%	52%	76%	62%	75%	86%	62%	50%
% funds diff. loading on SMB	89%	95%	84%	91%	87%	80%	80%	92%	63%	67%	91%	58%
% funds diff. loading on HML	57%	59%	74%	59%	70%	70%	89%	100%	71%	71%	76%	75%
Relative Performance												
% primary > secondary [of funds w/ secondary]	28.21%	-	86.36%	-	90.91%	-	23.08%	-	95.24%	-	83.33%	-
% primary > fund (1 BM)	29%	-	56%	-	36%	-	36%	-	43%	-	35%	-
% primary < fund (1 BM)	34%	-	15%	-	24%	-	36%	-	24%	-	32%	-
% primary > fund > Secondary (2 BM)	1%	-	3%	-	8%	-	1%	-	6%	-	0%	-
% secondary > fund > primary (2 BM)	5%	-	0%	-	1%	-	1%	-	0%	-	1%	-

	Small Growth		Mid-Cap Growth		Small Value		Fn. Large Growth		Mid-Cap Value		Mid-Cap Blend	
	Primary	Sec	Primary	Sec	Primary	Sec	Primary	Sec	Primary	Sec	Primary	Sec
% Funds w/ Benchmarks	99.35%	33.12%	100%	34.78%	100%	22.43%	97%	22%	100%	22.11%	98.89%	25.56%
# Unique Benchmarks	15	19	17	23	14	11	12	16	10	17	16	17
Benchmark Performance												
Avg. 3 year return	35.08%	31.76%	44.24%	39.61%	22.18%	22.5%	35.63%	32.33%	35.21%	29.73%	36.96%	31.11%
SD of 3 year return	12.63	11.74	10.67	14	14.17	6.3	4.96	8.43	10.61	7.22	10.22	11.71
Avg. volatility	4.54	4.49	3.85	4.06	4.68	4.47	3.24	3.28	3.91	3.96	3.89	4.04
Benchmark Fit												
Avg. R^2 of fund on benchmark	0.79	0.85	0.81	0.81	0.87	0.91	0.87	0.88	0.88	0.9	0.89	0.87
% funds diff. loading on market	81%	76%	76%	67%	67%	71%	69%	33%	63%	33%	74%	60%
% funds diff. loading on SMB	78%	71%	64%	71%	74%	93%	75%	100%	78%	67%	79%	90%
% funds diff. loading on HML	59%	88%	62%	67%	71%	71%	44%	50%	68%	44%	67%	60%
Relative Performance												
% primary > secondary [of funds w/ secondary]	47.06%	-	41.67%	-	100%	-	50%	-	88.89%	-	90%	-
% primary > fund (1 BM)	12%	-	24%	-	46%	-	7%	-	36%	-	39%	-
% primary < fund (1 BM)	47%	-	31%	-	25%	-	60%	-	31%	-	29%	-
% primary > fund > Secondary (2 BM)	1%	-	0%	-	5%	-	2%	-	1%	-	5%	-
% secondary > fund > primary (2 BM)	1%	-	5%	-	0%	-	1%	-	0%	-	0%	-

Note: All statistics using returns are calculated with 3 years of data from 2017 through 2019.

8. General Discussion

8.1 Summary of Findings

This study examined the use of benchmarks by mutual funds using a large and comprehensive dataset of funds in 12 sectors, and the reaction of individuals to various benchmark presentations in a large-scale experiment. Our results suggest wide variation in the way that funds use benchmarks and also that many individuals react quite strongly to different benchmark presentations. In our preliminary review of markets data, we document that:

- There are a relatively large number of benchmarks in use in each fund category, with some fund categories employing nearly two dozen primary benchmarks and over three dozen secondary benchmarks.
- Many funds (about 2/3 to 4/5 of funds in each sector we considered) did not choose to present a second benchmark.
- There is substantial variation in the performance of benchmarks that are employed within a sector. In particular, 10-year cumulative returns show performance return differentials among the benchmarks used within some sectors of over 400%. This variation makes it difficult to understand how reliable these benchmarks are as a reference point for fund performance.
- Some funds use extremely rare benchmarks (4.5%). Within the 12 sectors we reviewed, each sector tended to have between 2 and 13 benchmarks used by only one fund.
- We observed many different types of benchmark choices. For example, we found some examples of equity funds that use an equity index and a bond index as benchmarks. These observations highlight some of the variability in funds' benchmark choices.
- The definitions of broad and narrow benchmarks appear to be the subject of some interpretation. Although we do not assess the appropriateness of benchmark selection, we provide data that contextualizes benchmark appropriateness. The most common benchmark used is the S&P 500 Total Return Index, which about a quarter of funds select. In our data, only about half of funds present at least one benchmark that has a correlation with the S&P 500 Index of 0.95 or above.

Our qualitative research provided some interesting insights that set the stage for our quantitative experimental study, although these results are based on a small sample and not conclusive on their own. Our qualitative study provided preliminary evidence that:

- Investors may react to variations in the visual presentation of the relative position of funds and their benchmarks. This initial observation is difficult to contextualize in most economic models.
- Investors have different reference points for contextualizing fund fees.

Our experimental results built on these initial qualitative and market data observations and yielded extremely interesting conclusions. We developed a sophisticated, yet elegant research design that focused on a two-benchmark "sandwich case" in which one benchmark outperforms the fund and one benchmark is outperformed by the fund – one condition in which the broad benchmark outperformed the narrow and one where the narrow outperformed the broad benchmark. We created other conditions based on those two-benchmark conditions by removing one or both benchmark reference lines; a no-graph condition enabled us to understand the effect of benchmark presentation on expectations of future fund performance. Our primary outcomes of interest were subjective ratings of fund attractiveness and an incentivized investment participation outcome. Our design allowed us to study the role of the relative position of benchmarks, the number of benchmarks (zero, one or two), the benefits of an explanatory text defining the benchmarks that are used, and the relative impact of broad versus narrow benchmarks. We also were able to use our design to study the differential impact at different sophistication levels, the expectations formation process, and the effect of benchmark presentations on optimal allocation decisions.

Overall, there is substantial variation in the between-subjects responses for both outcomes of interest, both between and across conditions. Specifically, we observed:

- Fund attractiveness and incentivized allocation amounts were lower in the condition presenting a single benchmark above the fund. In the two benchmarks condition (one benchmark outperforming and one benchmark underperforming the fund) this effect was present, but more muted: a smaller decrease versus the baseline condition (graph, no benchmark), and a reduction in the statistical significance level of difference with the baseline condition (the incentivized allocation was not statistically different). The benchmark below the fund did not result in statistically different allocations or attractiveness vis-à-vis our baseline condition.
- Although many respondents gave survey responses that suggested they were inclined to regard the benchmarks as marketing devices selected in order to show the fund in a valuable light, rather than as a decision-viable reference tool, we do not find evidence that participants entirely disregarded benchmarks. At the same time, participants reacted in their attractiveness ratings most strongly (negatively) to a benchmark presentation where at least one benchmark outperformed the fund, and individuals indicated a higher interest in searching for alternatives to the Middlewood Materials Fund when a single benchmark outperformed the fund.
- We observe that sophistication matters for participants' reactions to benchmarks, but it matters in a way that is quite different than most economic models would assume and much of regulatory theory seems to be grounded on. Our results suggest that the most sophisticated participants were more reactive to benchmark presentations than lower sophistication participants.

- We do not find evidence supporting the notion that participants believed that the narrow benchmark is a better reference point than the broad benchmark. In fact, participants in our study did not react differently to the broad and narrow benchmarks.
- We did not find evidence that the textual clarifications of benchmarks improved investor comprehension or altered fund attractiveness ratings or participation decisions. They also did not alter the (non-)distinction that study participants made between broad and narrow benchmarks.
- The no graph condition had substantially lower ratings of attractiveness and lower investment rates in our allocation task. This is not surprising in our experimental context because we provided very little information to participants. This condition was not so much added as a control condition, but rather as a way to better understand if and how benchmark presentations affected expectations formation.
- In our setup, expectations would seem to provide a key role in many economic frameworks in how benchmark presentation affects incentivized decisions and, to a lesser extent, attractiveness ratings. Expectations of future fund performance varied slightly by experimental condition, in contrast with most standard economic models. The effect is muted when comparing to our baseline condition (graph with no benchmarks), but the single benchmark above the fund and single benchmark below the fund conditions do result in statistically significantly different expectations of future fund performance.
- We used a simple economic model to assess the extent to which a particular condition resulted in a deviation from the expected utility-maximizing allocation in our incentivized allocation task. We observed that, overall, two conditions resulted in a distortion from the optimal allocation. These were the single benchmark above (outperforming) the fund, and the single benchmark below (underperforming) the fund. These conditions led to a respective increase (benchmark below) or decrease (benchmark above) in expected future returns for the fund, which mechanically altered the optimal allocation in each of these conditions, but overall led to a distortion in which a lower than optimal amount was allocated in both cases by a statistically significant amount close to $800. Because of the increase (decrease) in expected returns in the single benchmark below (above) condition, our simple model increased (decreased) the required investment amount for utility maximization; in the end, we observed that both conditions distorted allocation from the utility maximization allocation in a similar amount. Of course, the fact that these conditions changed expectations might itself be a source of welfare loss. Our examination of these results by subgroup suggests that much of this is driven by non-investors, but also that independent investors' optimal allocations are also at least marginally affected by some of the single benchmark conditions.

In Section 7 we returned to the markets data to provide additional context to the experiment and the earlier market results. Our work provided new insights into the relationship of funds and their benchmarks and context to the argument that the narrow, or secondary, benchmark is a better benchmark than the broad-based market benchmark that funds are required to use. We observed that:

- There is very little support in our analysis for the claim that secondary benchmarks currently used by funds provide a more relevant comparison for investors than primary benchmarks that funds use (recall that, in our analyses, we analyzed funds with two benchmarks and classified benchmarks as "secondary" when they had a lower correlation to the S&P 500 Index). We examined the markets data in two ways, with the goal of understanding the "fit" between fund performance and benchmark performance. In particular, our first analysis compared the average R^2 from the a simple regression of fund performance on the primary benchmark and a separate regression on the secondary benchmark. The average values differ slightly in some sectors, but overall we do not find that the secondary benchmark is a better fit than the primary benchmark in the sense that more variance in fund performance was explained. In fact, in most cases the secondary fits the fund's performance about as well as the primary benchmark, and actually tends to fit worse than the primary in the majority of sectors we examined. Our second analysis fit a Fama-French factor model to determine whether there was significantly different fit in factor loadings. These estimates exhibited some differences between the primary and the secondary benchmarks, but did not lead to a consistent observation that the secondary benchmarks are a better fit to fund performance than the primary benchmarks.
- The sandwich case positioning of benchmarks and funds in our experimental conditions may appear a special situation, but in reality, the experimental conditions we created represent a large fraction of presentation conditions experienced in the wild. In each of the 12 sectors we studied, our presentation cases represented as few as 60 percent of funds in the sector and as many of 76 percent of funds in the sector.

8.2 Limitations

Our work, as any research, is not without limitations. Perhaps the biggest limitation is that, for our particular experiment, we had to choose stimuli that were able to be digested by participants rapidly and that would reflect theoretically interesting variation. Future work may extend our results by studying a broader range of benchmarks and performance histories for different hypothetical funds, as well as alternatives to the 10-year cumulative performance line graph that we examined.

8.3 Conclusion

Considerable research suggests that investors prioritize information on investment performance and use it to make decisions that may affect their ability to meet their financial goals and achieve financial well-being. As such, understanding reactions to performance information, and comparative benchmark information that is required to accompany performance disclosures, is critical. Using a novel, large-scale experiment with a national sample, as well as in-depth analysis of real-world benchmark use, we have presented a comprehensive set of findings on how funds use benchmarks and how investors may react to them.

More broadly, past work has argued that mandatory disclosures should be tested with consumers to ensure that communication objectives (e.g., awareness, comprehension) are achieved (Kozup et al., 2012). As such, we contribute to debates that raise questions about consumers' knowledge of financial products and what consumers can learn from disclosures (e.g., CFPB, 2020; Chin and Bruine de Bruin, 2019; Chin, Scholl, and VanEpps, 2021; Hogarth and Merry, 2011; Kleimann, 2013; Lacko and Pappalardo, 2010; Scholl, Craig, and Chin, 2022). We hope these findings are used to better understand investor decision-making processes, support investor protection efforts, and welfare.

References

Athey, Susan, and Guido W. Imbens. "The econometrics of randomized experiments." In *Handbook of economic field experiments*, vol. 1, pp. 73-140. North-Holland, 2017.

Armantier, Olivier, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. "An overview of the survey of consumer expectations." *Economic Policy Review* 23-2 (2017): 51-72.

Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. "The price is right: Updating inflation expectations in a randomized price information experiment." *Review of Economics and Statistics* 98, no. 3 (2016): 503-523.

Armona, Luis, Andreas Fuster, and Basit Zafar. "Home price expectations and behavior: Evidence from a randomized information experiment." *Staff Report, No. 798, Federal Reserve Bank of New York* (2016).

Barber, Brad M., Terrance Odean, and Lu Zheng. "Out of sight, out of mind: The effects of expenses on mutual fund flows." *The Journal of Business* 78, no. 6 (2005): 2095-2120.

Barberis, Nicholas, Lawrence J. Jin, and Baolian Wang. "Prospect theory and stock market anomalies." The Journal of Finance 76, no. 5 (2021): 2639-2687.

Beneish, Messod D., and Robert E. Whaley. "A scorecard from the S&P game." Journal of Portfolio Management 23, no. 2 (1997): 16.

Ben-Shahar, Omri and Carl E. Schneider. "The failure of mandated disclosure." *University of Pennsylvania Law Review* 159, no. 3 (2011): 647–749.

Bruine de Bruin, Wändi, Alycia Chin, Jeff Dominitz, and Wilbert van der Klauuw, "Household surveys and probabilistic questions" in *Handbook of Economic Expectations*, ed. Ruediger Bachmann (Elsevier, 2022).

Chin, Alycia, and Wändi Bruine de Bruin. "Helping consumers to evaluate annual percentage rates (APR) on credit cards." *Journal of Experimental Psychology: Applied* 25, no. 1 (2019): 77.

Chin, Alycia, Brian Scholl, and Eric M. VanEpps. "Jargon in fund fee disclosures." Office of the Investor Advocate Working Paper, *Washington, DC: Office of the investor Advocate*. (2021).

Chin, Alycia, David Zimmerman, Heidi Johnson, and Suzanne B. Shu. "Disclosure Design, Consumer Comprehension, and Decisions about Overdraft Services." *Working paper*. (2022).

Choi, James J., David Laibson, and Brigitte C. Madrian. "Why does the law of one price fail? An experiment on index mutual funds." *The Review of Financial Studies* 23, no. 4 (2010): 1405-1432.

Choi, James J., and Adriana Z. Robertson. "What Matters to Individual Investors? Evidence from the Horse's Mouth." *The Journal of Finance* LXXV, no. 4 (2020): 1965-2020. doi: 10.1111/jofi.12895

Consumer Financial Protection Bureau (CFPB). "Disclosure of Time-Barred Debt and Revival." Retrieved at: https://files.consumerfinance.gov/f/documents/cfpb_debt-collection-quantitative-disclosure-testing_report.pdf (2020).

Cremers, K.J. Martijn, Jon A. Fulkerson, and Timothy B. Riley. "Benchmark discrepancies and mutual fund performance evaluation." *Journal of Financial and Quantitative Analysis* 57, no. 2 (2022): 543-571.

Cremers, K.J. Martijn, and Antti Petajisto. "How active is your fund manager? A new measure that predicts performance." *The Review of Financial Studies* 22, no. 9 (2009): 3329-3365.

Eckel, Catherine C., and Philip J. Grossman. "Sex differences and statistical stereotyping in attitudes toward financial risk." *Evolution and Human Behavior* 23, no. 4 (2002): 281-295.

Egan, Mark. "Brokers versus retail investors: Conflicting interests and dominated products." *The Journal of Finance* 74, no. 3 (2019): 1217-1260.

Engelberg, Joseph, Charles F. Manski, and Jared Williams. "Comparing the point predictions and subjective probability distributions of professional forecasters." *Journal of Business & Economic Statistics* 27, no. 1 (2009): 30-41.

Fidelity Investments, Comment letter on Tailored Shareholder Reports, Treatment of Annual Prospectus Updates for Existing Investors, and Improved Fee and Risk Disclosure for Mutual Funds and Exchange-Traded Funds; Fee Information in Investment Company: File Number S7-09-20 (January 4, 2021) https://www.sec.gov/comments/s7-09-20/s70920-8204333-227469.pdf

Fisch, Jill E. and Tess Wilkinson-Ryan. "Why do retail investors make costly mistakes? An experiment on mutual fund choice." *University of Pennsylvania Law Review*, *162*, no. 3 (2014): 605-647.

Freedman, David A. *Statistical models: theory and practice*. Cambridge University Press, 2009.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. "Five facts about beliefs and portfolios." American Economic Review 111, no. 5 (2021): 1481-1522.

Hogarth, Jeanne M., and Ellen A. Merry. "Designing disclosures to inform consumer financial decisionmaking: Lessons learned from consumer testing." *Federal Reserve Bulletin* 97, no. August (2011).

Hortaçsu, Ali, and Chad Syverson. "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds." *The Quarterly Journal of Economics* 119, no. 2 (2004): 403-456.

Hsee, Christopher K. "The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives." *Organizational Behavior and Human Decision Processes* 67, no. 3 (1996): 247-257.

Hsee, Christopher K., and Jiao Zhang. "General evaluability theory." *Perspectives on Psychological Science* 5, no. 4 (2010): 343-355.

Investment Company Institute, Comment letter on the SEC Proposal on Tailored Shareholder Reports, Treatment of Annual Prospectus Updates for Existing Investors, and Improved Fee and Risk Disclosure for Mutual Funds and Exchange-Traded Funds; Fee Information in Investment Company Advertisements (Dec. 21, 2020). Retrieved from: https://www.sec.gov/comments/s7-09-20/s70920-8186011-227164.pdf

Investment Company Institute (ICI). "2021 Investment Company Fact Book." (2021a). Retrieved from: https://www.ici.org/system/files/2021-05/2021_factbook.pdf

Investment Company Institute (ICI). "What US Households Consider When They Select Mutual Funds, 2020." *ICI Research Perspective* 27, no. 4 (2021b): 1-12.

John Hancock Investment Management LLC, Comment letter on Tailored Shareholder Reports, Treatment of Annual Prospectus Updates for Existing Investors, and Improved Fee and Risk Disclosure for Mutual Funds and Exchange-Traded Funds; Fee Information in Investment Company Advertisements (File No. S7-09-20) (January 4, 2021). Retrieved from: https://www.sec.gov/comments/s7-09-20/s70920-8204305-227456.pdf

Johnson, Joseph M., Gerard J. Tellis, and Noah VanBergen. "Fooled by success: how, why, and when disclosures fail or work in mutual fund ads." *Journal of Public Policy & Marketing* 41, no. 1 (2022): 54-71.

Kleimann Communication Group, Inc. “Know Before You Owe: Quantitative Study of the Current and Integrated TILA-RESPA Disclosures.” (2013). Retrieved from: https://files.consumerfinance.gov/f/201311_cfpb_study_tila-respa_disclosure-comparison.pdf

Kozup, John, Elizabeth Howlett, and Michael Pagano. “The Effects of Summary Information on Consumer Perceptions of Mutual Fund Characteristics.” *Journal of Consumer Affairs* 42, no. 1 (2008): 37-59.

Kozup, John, Charles R. Taylor, Michael L. Capella, and Jeremy Kees. “Sound disclosures: Assessing when a disclosure is worthwhile.” *Journal of Public Policy & Marketing* 31, no. 2 (2012): 313-322.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. “How elastic are preferences for redistribution? Evidence from randomized survey experiments.” *American Economic Review* 105, no. 4 (2015): 1478-1508.

Lacko, James M., and Janis K. Pappalardo. “The failure and promise of mandated consumer mortgage disclosures: Evidence from qualitative interviews and a controlled experiment with mortgage borrowers.” *American Economic Review* 100, no. 2 (2010): 516-21.

Latham, Scott, and Michael Braun. "Does short-termism influence firm innovation? An examination of S&P 500 firms, 1990-2003." Journal of Managerial Issues (2010): 368-382.

Larrick, Richard P., Jack B. Soll, and Ralph L. Keeney. “Designing better energy metrics for consumers.” *Behavioral Science & Policy* 1, no. 1 (2015): 63-75.

Markowitz, H. “Portfolio Selection.” *The Journal of Finance* 7, no. 1 (1952): 77–91.

Morgan, M. Granger, Baruch Fischhoff, Ann Bostrom, and Cynthia J. Atman. *Risk communication: A mental models approach*. Cambridge University Press, 2001.

Mullally, Kevin, and Andrea Rossi. “Benchmark Backdating in Mutual Funds.” *Available at SSRN 3887838* (2021).

Pavlova, Anna, and Taisiya Sikorskaya. “Benchmarking intensity.*” Available at SSRN 3689959* (2022).

Pontari, Beth A., Andrea J.S. Stanaland, and Tom Smythe. "Regulating information disclosure in mutual fund advertising in the United States: Will consumers utilize cost information?" *Journal of Consumer Policy* 32, no. 4 (2009): 333-351.

Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei. "Marketing mutual funds." *The Review of Financial Studies* 34, no. 6 (2021): 3045-3094.

Scholl, Brian, Adam W. Craig, and Alycia Chin. "Helping People Make Decisions about Mutual Funds using Visual Aids," *Office of the Investor Advocate Working Paper 2022-01*.

Scholl, Brian, and Angela Fontes. "Measuring Public Knowledge of Mutual Funds," *Office of the Investor Advocate Working Paper 2021-22*. (2021).

Scholl, Brian, and Angela Fontes. "Mutual fund knowledge assessment for policy and decision problems." *Financial Services Review* 30, no. 1 (2022): 31-56.

Scholl, Brian, Dan Silverman and Marco Enriquez (2021), "Disclosure Complexity and Fund Performance," Office of the Investor Advocate Working Paper, forthcoming.

Sensoy, Berk A. "Performance evaluation and self-designated benchmark indexes in the mutual fund industry." *Journal of Financial Economics* 92, no. 1 (2009): 25-39.

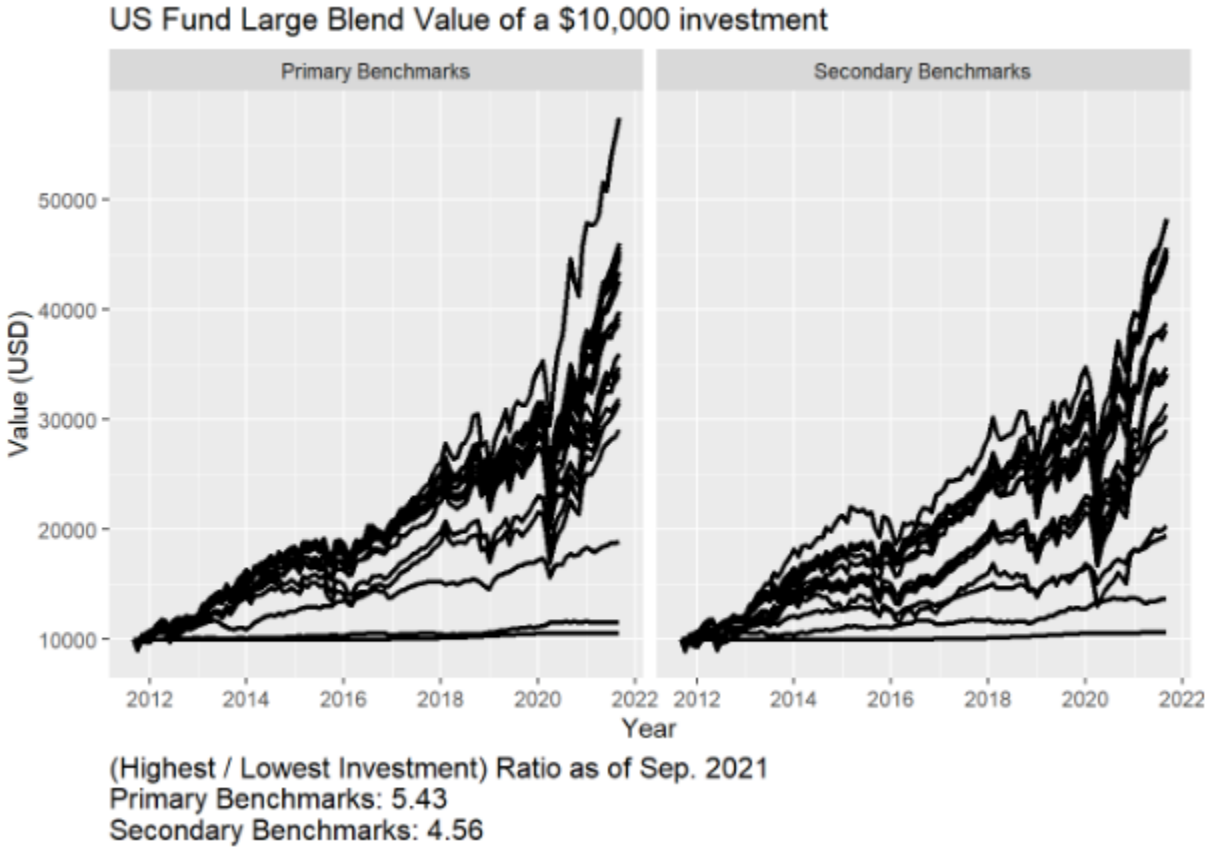
Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *The Journal of Finance* 19, no. 3 (1964): 425–442.

Thorp, Susan, Hazel Bateman, Loretti I. Dobrescu, Ben R. Newell, and Andreas Ortmann. "Flicking the switch: Simplifying disclosure to improve retirement plan choices." *Journal of Banking & Finance* 121 (2020): 105955.

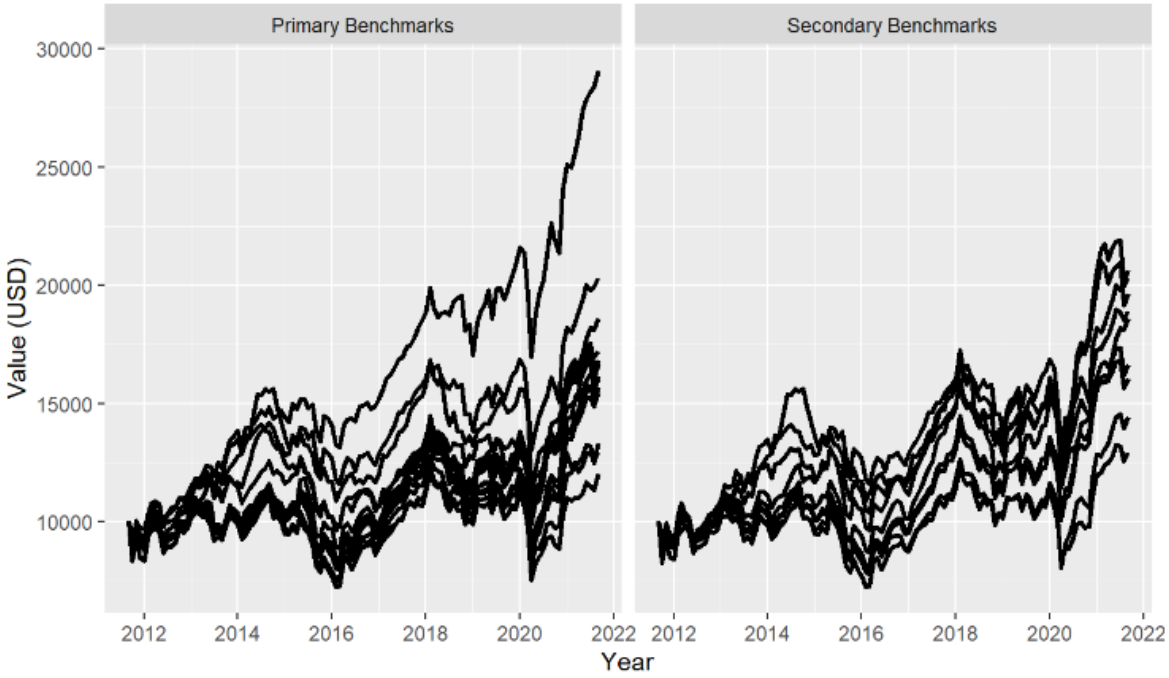
Appendices

Appendix A. Additional Figures on Performance Variation

This section contains additional spaghetti plots showing variation in benchmark performance over a 10-year period. These figures are analogous to Figure XXX in the manuscript.

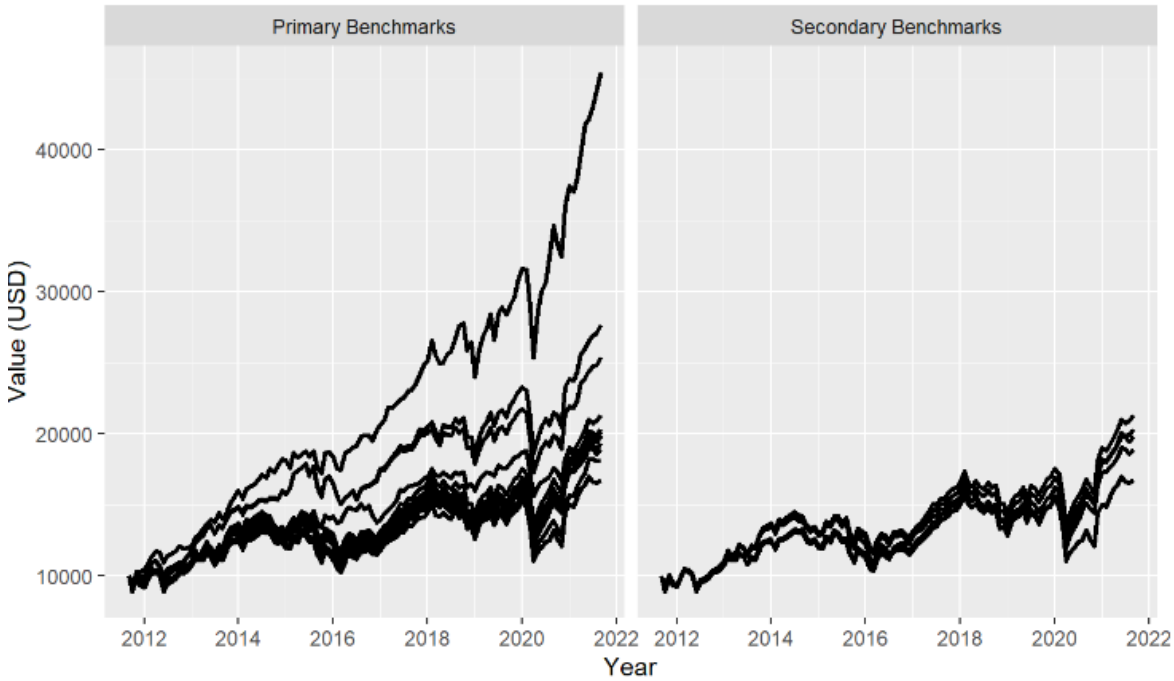


US Fund Diversified Emerging Mkts Value of a $10,000 investment



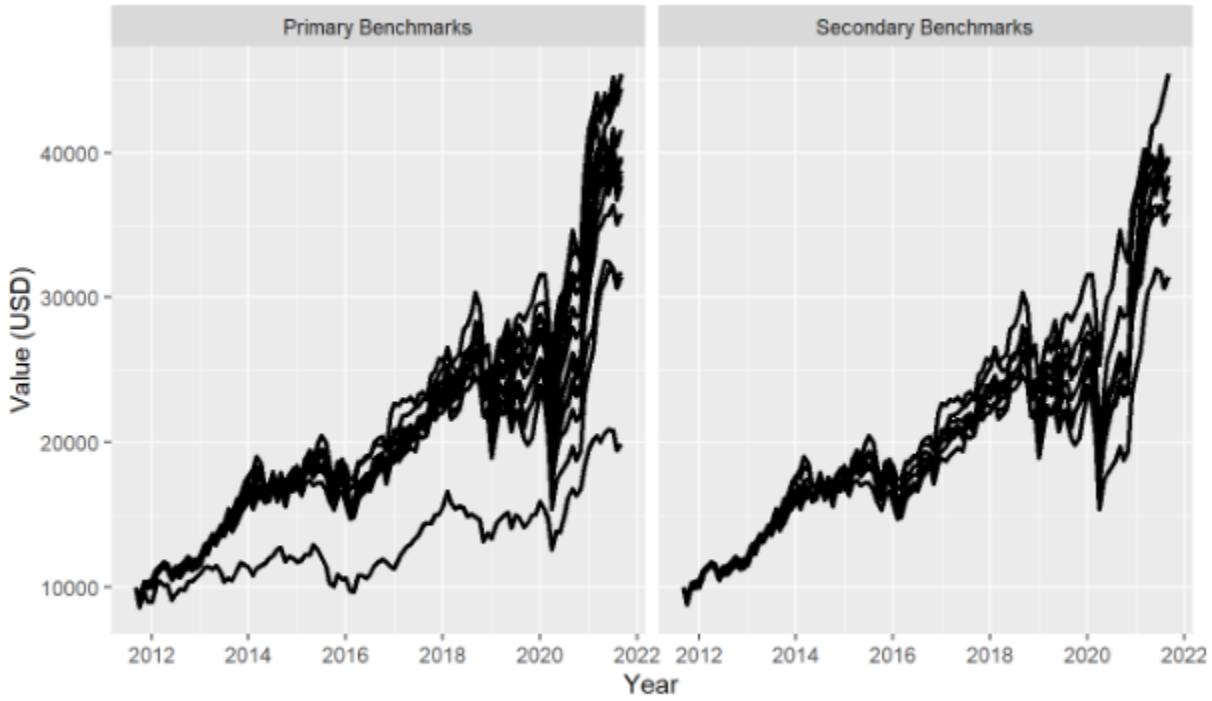
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 2.42
Secondary Benchmarks: 1.6

US Fund Foreign Large Blend Value of a $10,000 investment



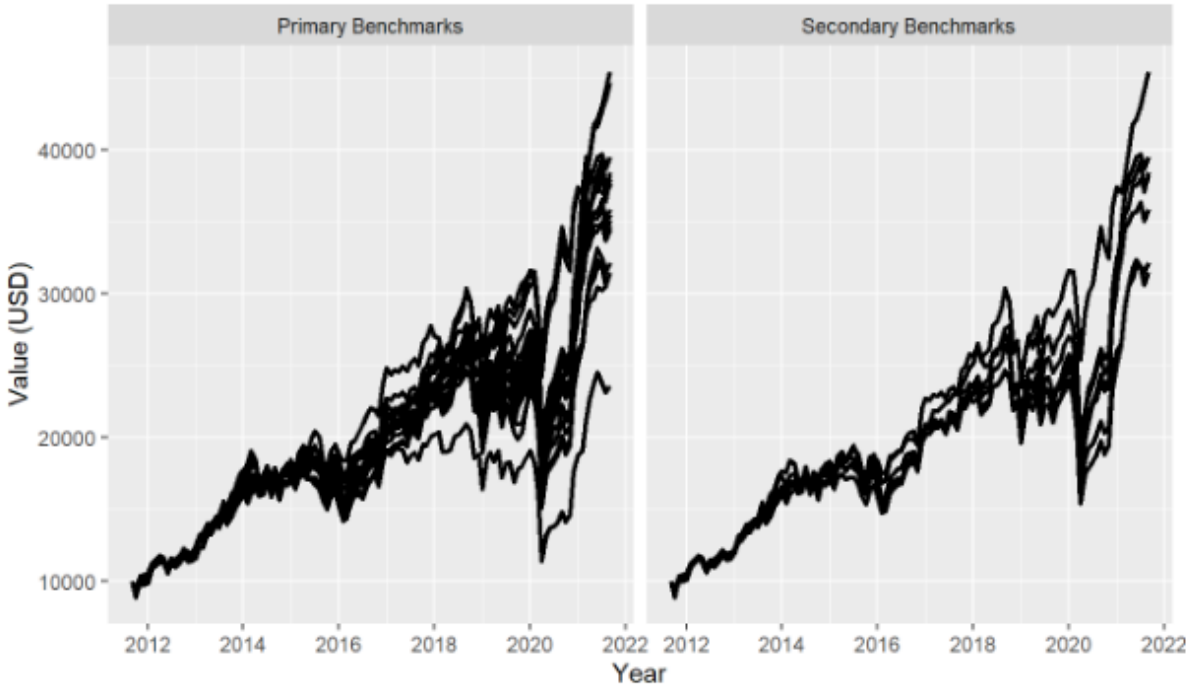
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 2.72
Secondary Benchmarks: 1.27

US Fund Small Growth Value of a $10,000 investment



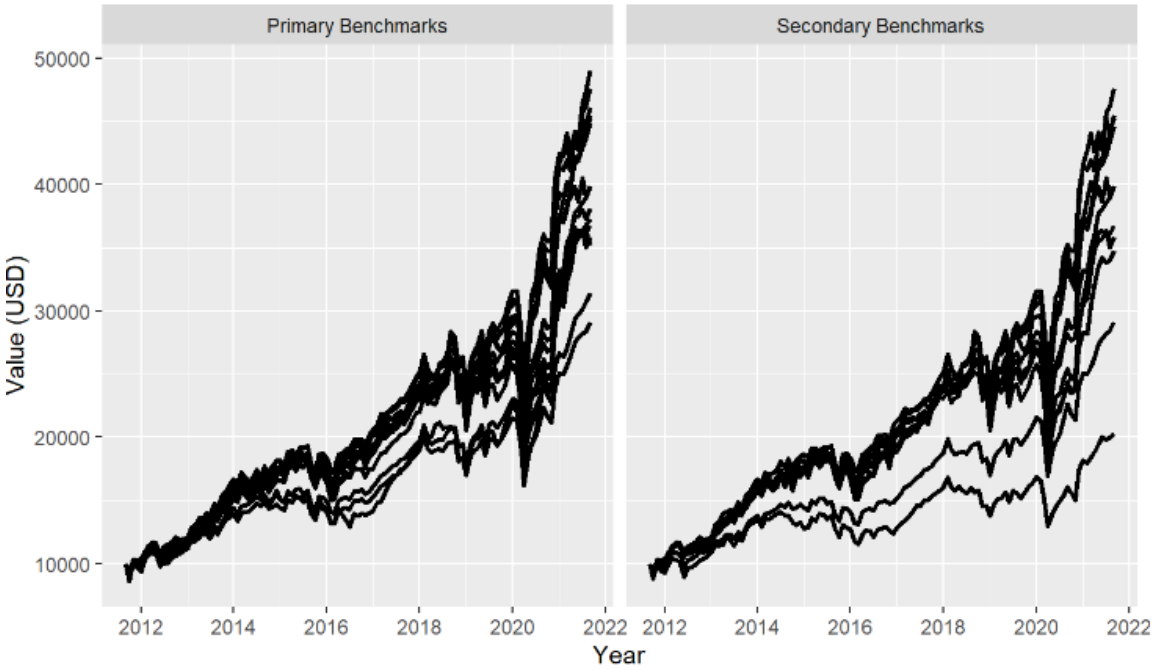
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 2.28
Secondary Benchmarks: 1.44

US Fund Small Blend Value of a $10,000 investment



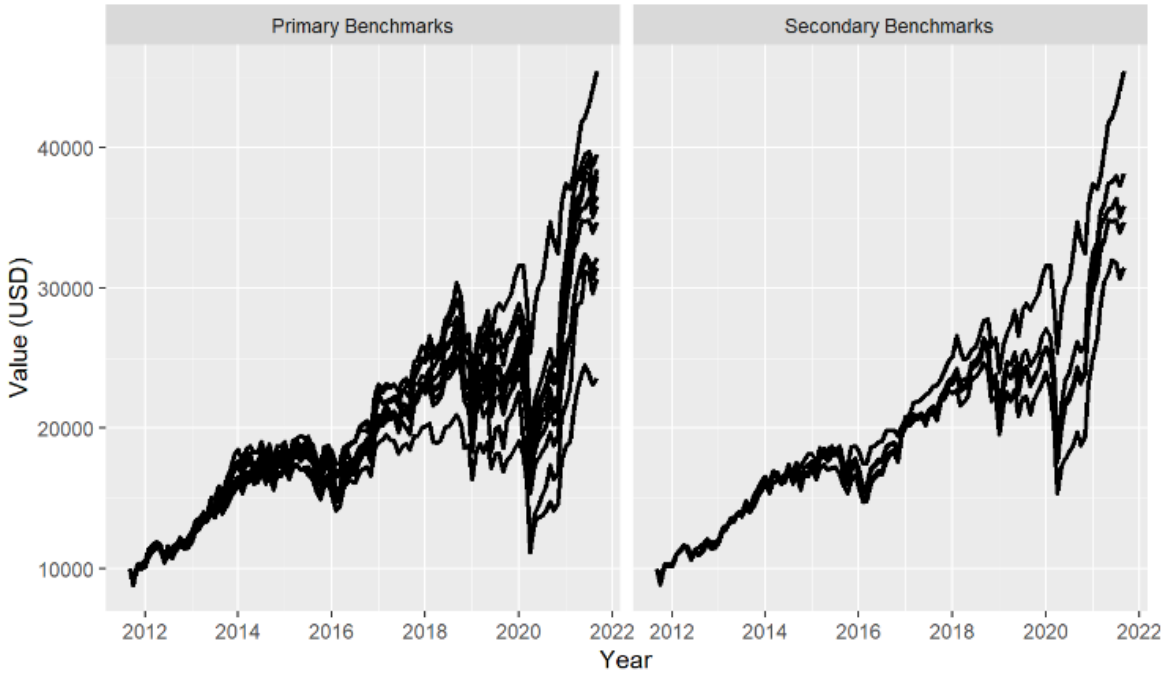
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.93
Secondary Benchmarks: 1.44

US Fund Mid-Cap Growth Value of a $10,000 investment



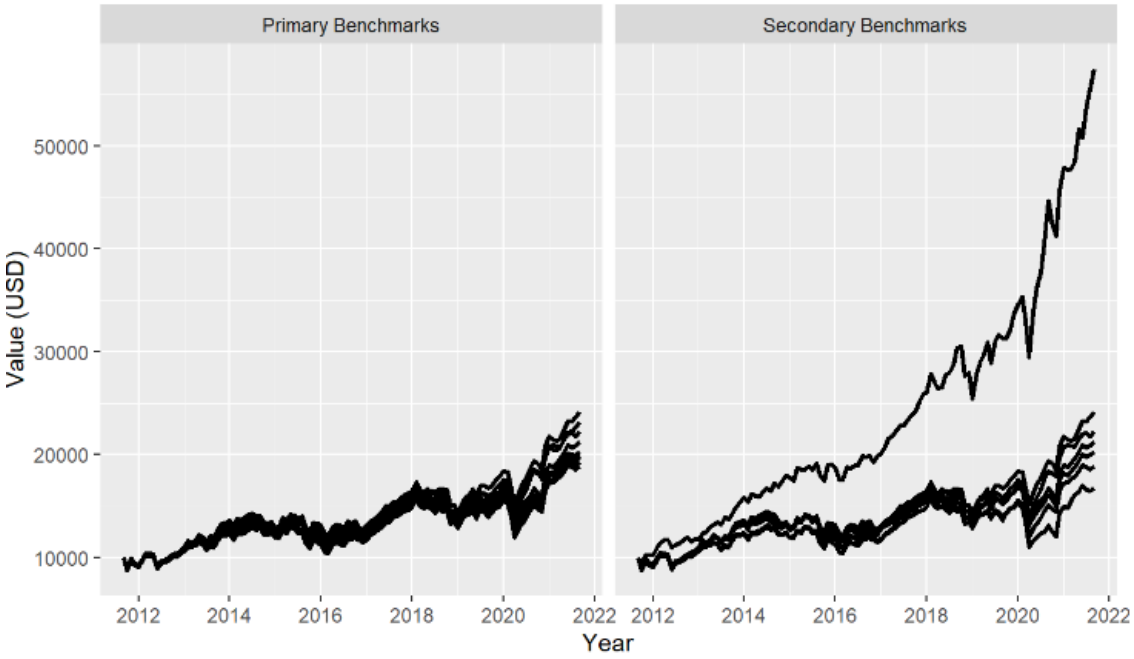
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.69
Secondary Benchmarks: 2.35

US Fund Small Value Value of a $10,000 investment



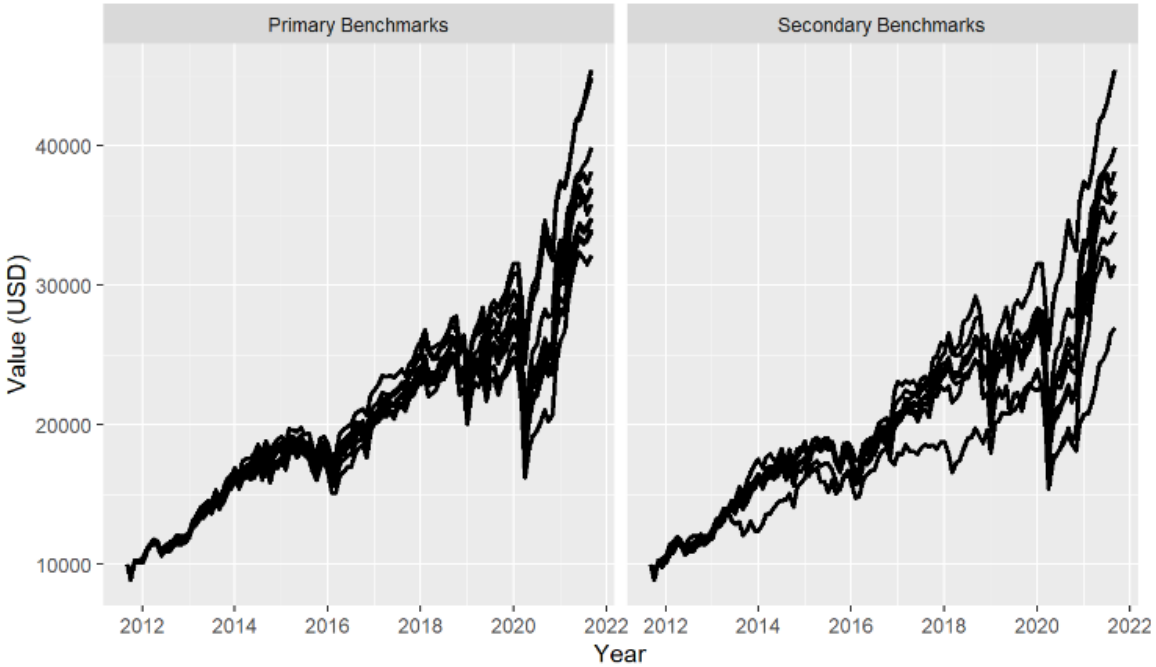
(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.93
Secondary Benchmarks: 1.44

US Fund Foreign Large Growth Value of a $10,000 investment

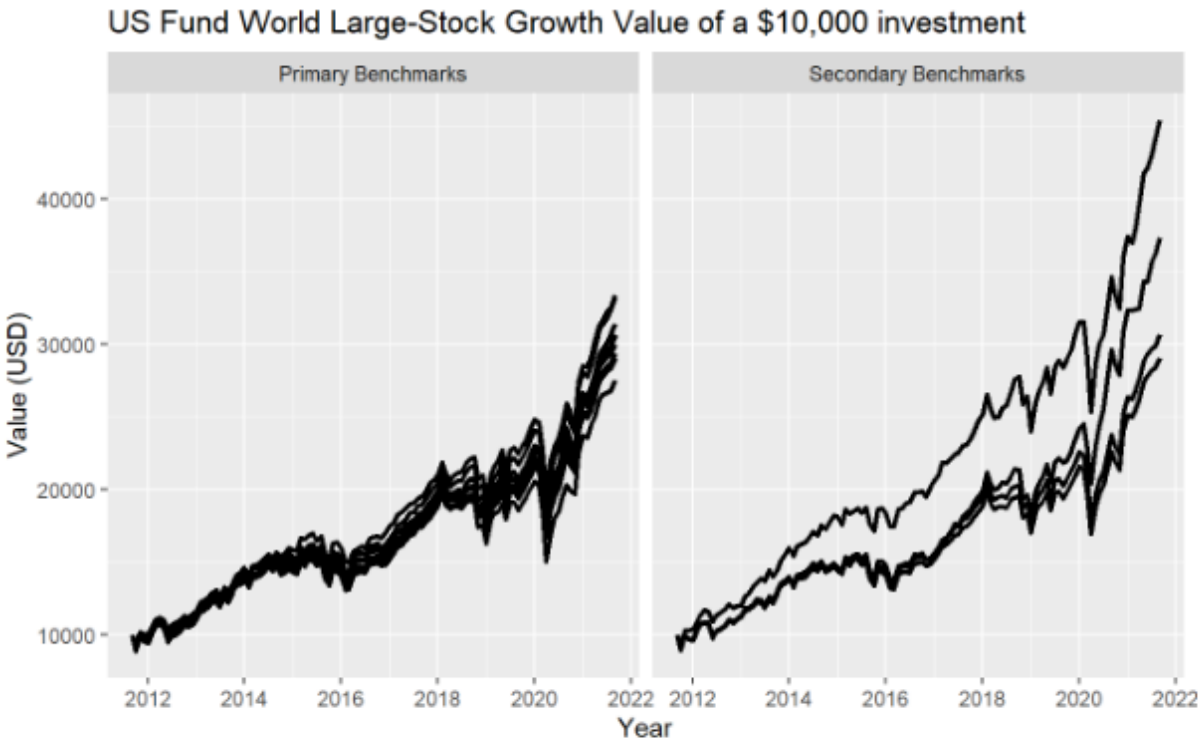


(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.28
Secondary Benchmarks: 3.44

US Fund Mid-Cap Value Value of a $10,000 investment



(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.41
Secondary Benchmarks: 1.68



(Highest / Lowest Investment) Ratio as of Sep. 2021
Primary Benchmarks: 1.21
Secondary Benchmarks: 1.56

Appendix B. Additional Information on Qualitative Pilot

This appendix contains additional detail on the 16 qualitative interviews that we conducted as part of the pilot testing for the testing described in this report. The primary purpose of the interviews was to gather preliminary information and generate ideas that would inform subsequent quantitative testing. We asked interview participants to comment on a mock-up of a fund's annual shareholders report, point out areas of interest and confusion, and react to information we presented.

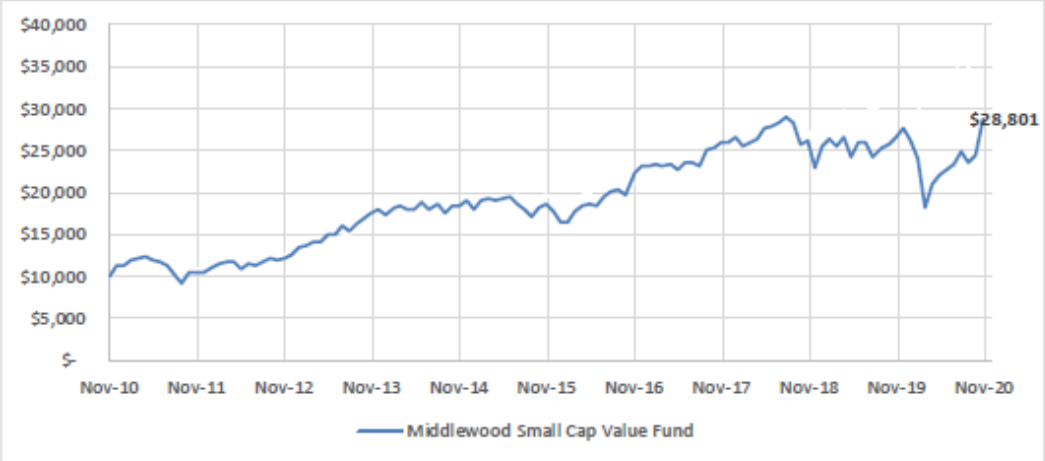
As described in the manuscript, we showed participants four performance graphs. The figure immediately below (Figure A1) shows the first performance graph used in the qualitative interviews, whereas the second figure (Figure A2) shows the fourth and final performance graph used in the qualitative interviews, including the accompanying narrative text explaining the meaning of the benchmark lines.

Figure A1. Initial performance graph shown in qualitative interviews.

How did the Fund perform over the past 10 years?

Keep in mind that the Fund's past performance is not a good predictor of how the Fund will perform in the future.

Cumulative Performance: December 1, 2010 through November 30, 2020
Initial Investment of $10,000

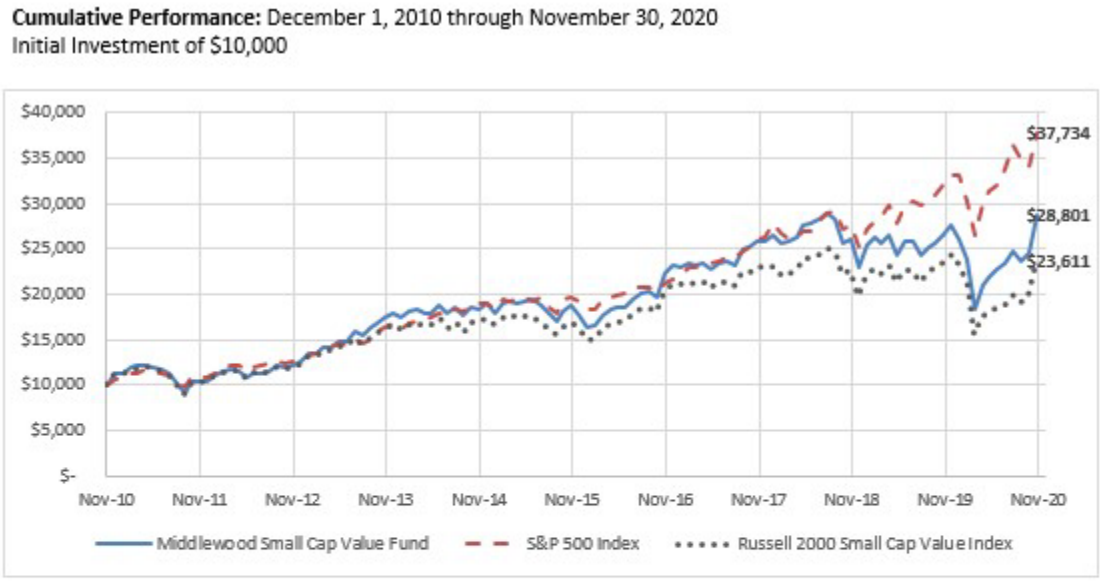


Average Annual Total Returns:

	1 Year	5 Years	10 Years
Class A	8.61%	9.01%	11.16%
Class Z	9.04%	9.44%	11.59%

Visit *www.middlewood.com/MSCV* or the Middlewood app for more recent performance information.

Figure A2. Final performance graph shown in qualitative interviews.



The performance graph compares the Middlewood Small Cap Value Fund to two indexes.

The first index reflects the broader stock market. This index allows you to see whether it could be valuable to pick investments in the stock market as a whole, instead of the investments that the fund chose. In this case, the first index is the S&P 500, which is generally considered representative of the overall U.S. stock market.

The second index consists of investments that are matched to the fund in terms of risk and strategy, which allows you to see how the fund is performing relative to similar investments. In this case, the second index is the Russell 2000 Value Index, an index of small public U.S. companies that are thought to be undervalued by the market.

As noted above, the interviews were designed for idea generation; with a small sample of 16 respondents, any conclusions are necessarily tentative and preliminary, and would benefit from follow-up testing with a larger sample (a methodology recommended in, for example, Morgan et al. 2001). In particular, based on our analysis of the interviews, the research team recommends additional research devoted to the following potential issues:

1. **Subjective evaluations of funds' cost**: Participants' impressions of the relative cost of a fund varied. Some participants had ways of judging expenses that make them vulnerable to overpaying. For instance, some participants reported rules of thumb that referenced past jobs in sales, or discrete fee cutoffs.

2. **Mutual fund share classes**: Multiple participants stated they had "no idea" or "did not know" what share classes meant (among others, Male, age 44 and Female, age 66). To the extent that share classes are a necessary component of other disclosures, future research should explore ways of explaining share classes to investors.

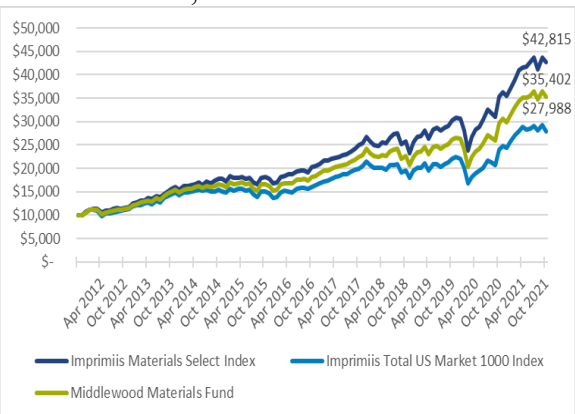
3. **Clarifying the role of the shareholder report**: Some participants expressed a lack of understanding of what to do with some of the information and how to use it most effectively for decision-making.

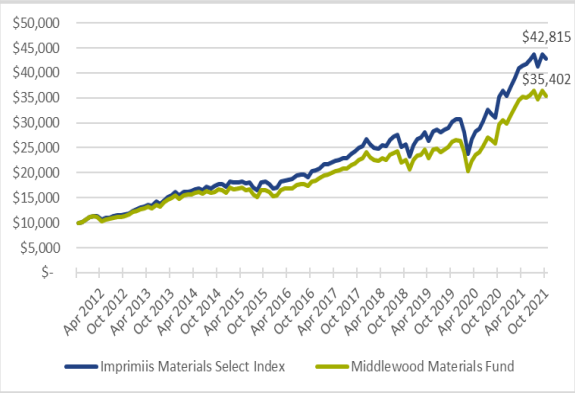
Appendix C. Additional Detail on Experimental Stimuli

The stimuli for the experiment were generated from actual indexes. We identified broad-based and industrial indexes with similar average annual returns. One pair of broad-based and industrial indexes had an average monthly return of around 1.3% and another pair had an average monthly return of just under 1%. We averaged the pair with the higher monthly return to create our high benchmark and average the lower pair to create the low benchmark. We chose to average these indexes so that the resulting indexes could plausibly serve as both a narrow index for the industrial sector and a broad-based index. We created the Middlewood Materials Fund by averaging the high and low benchmarks, to ensure that it would be in between them.

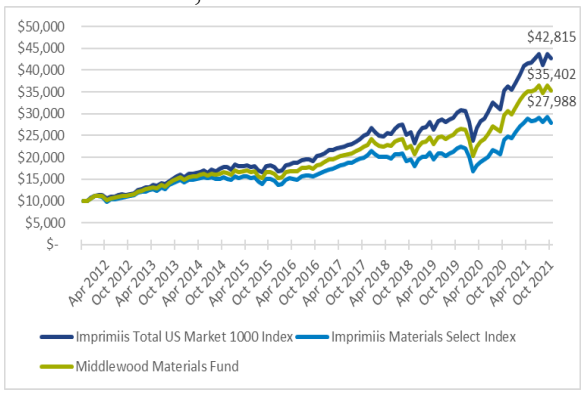
Full set of graphs shown

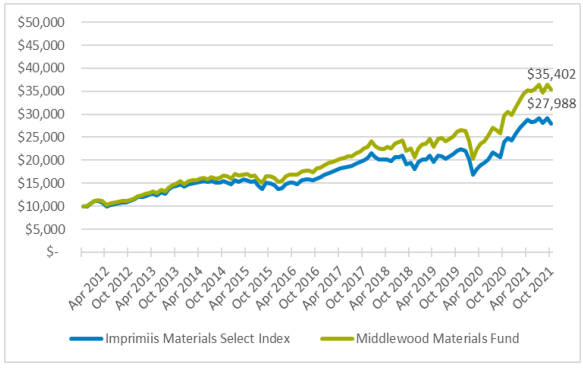
Narrow Above, Broad Below Conditions

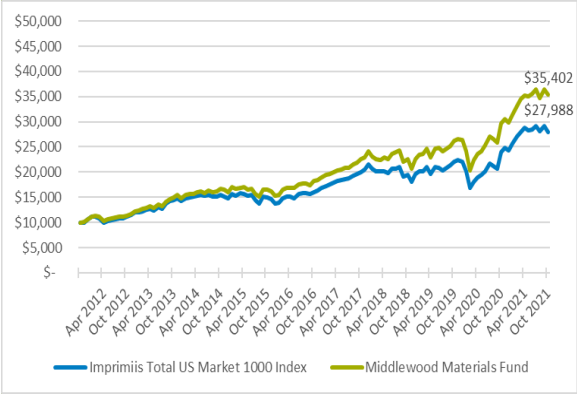


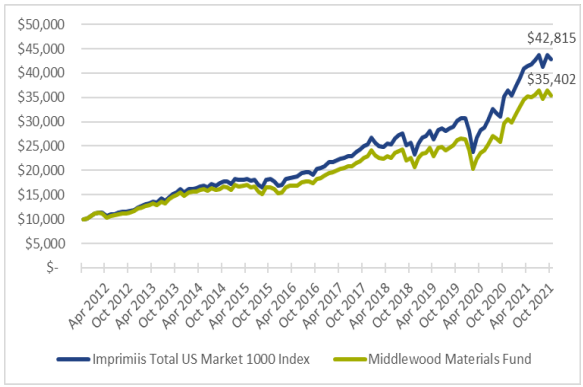


Narrow Below, Broad Above Conditions









No Benchmark Conditions

(No graph condition does not display a graph)

Appendix D. Assignment to Treatment

We examine assignment to treatment conditional on finishing the survey, following the procedure described in Kuziemko, Norton, Saez, and Stantcheva (2015). Specifically, we estimate multinomial logit regressions of the form:

$$P(Treatment = Treatment_i) = \frac{e^{\alpha_i + \beta_i Covariate + \varepsilon_i}}{\sum_{j \in Treatments} e^{\alpha_j + \beta_j Covariate + \varepsilon_j}}$$

where *Covariate* represents one of the variables shown in the table, and *i* and *j* represent one of the eight treatments (Condition 1 is our base treatment, so a coefficient is not estimated for Condition 1). Across the coefficients, only 4 had a p-value significant at the 5% level – that is to say, 5.7% (= 4/70) of coefficients were significant at the 5% level. Therefore, we need not be very concerned that the 25 covariates we consider correlate with assignment to treatment conditional on finishing the survey.

Table D.1. Ability of covariates to predict treatment condition

	p-values for condition						
Variable	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age	0.327	0.0658	0.672	0.0724	0.0925	0.236	0.266
White Non-Hispanic	0.727	0.519	0.767	0.203	0.306	0.0484	0.206
Black Non-Hispanic	0.974	0.0475	0.196	0.0809	0.0152	0.050	0.0862
Other Non-Hispanic	0.333	0.345	0.105	0.239	0.723	0.901	0.514
Hispanic	0.986	0.628	0.106	0.182	0.490	0.418	0.987
Two or More Races	0.787	0.736	0.0858	0.249	0.885	0.854	0.631
Male	0.633	0.924	0.715	0.947	0.359	0.222	0.446
Income in $1000s (based from bin midpoints)	0.0952	0.189	0.460	0.780	0.411	0.195	0.438
Mutual Fund Literacy score	0.182	0.886	0.451	0.840	0.701	0.448	0.806
Eckel-Grossman Lottery Choice	0.464	0.340	0.110	0.357	0.913	0.227	0.493

Appendix E. Supplementary Regression Tables

Table E.1. Regressions of fund attractiveness and allocations to the fund

	Attractiveness	Allocation
	(1)	(2)
No Graph	-19.167***	-1618.667***
	(1.667)	(275.457)
Single benchmark above fund	-8.574***	-779.103***
	(1.474)	(230.886)
Single benchmark below fund	0.220	-87.487
	(1.426)	(226.954)
Two benchmarks	-4.781**	-375.079
	(1.921)	(322.007)
Any narrative	0.045	414.313
	(1.838)	(313.927)
Guaranteed Return of 4%		-1341.813***
		(44.847)
Guaranteed Return of 6%		-2512.655***
		(63.700)
Constant	65.036***	10992.060***
	(1.165)	(186.073)
Observations	4,226	12,434
R2	0.047	0.054
Adjusted R2	0.046	0.054

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors are used for the evaluation regression and standard errors for the allocation regressions are clustered at the participant level.

Table E.2. Regressions of fund attractiveness and allocations to the fund using all eight experimental conditions

	Attractiveness (1)	Allocation (2)
No graph	-19.167***	-1,618.666***
	(1.668)	(275.491)
Single benchmark above fund (broad)	-8.044***	-669.426**
	(1.750)	(275.811)
Single benchmark above fund (narrow)	-9.074***	-881.089***
	(1.710)	(264.421)
Single benchmark below fund (broad)	0.009	-30.150
	(1.629)	(262.498)
Single benchmark below fund (narrow)	0.433	-145.668
	(1.665)	(263.764)
Two benchmarks with narrow above	-3.272	-449.084
	(2.106)	(352.495)
Two benchmarks with broad above	-6.270***	-303.052
	(2.091)	(351.771)
Any narrative	0.038	415.449
	(1.835)	(313.789)
Guaranteed return of 4%		-1,341.841***
		(44.851)
Guaranteed return of 6%		-2,512.539***
		(63.706)
Constant	65.036***	10,992.026***
	(1.165)	(186.095)
Observations	4,226	12,434
R^2	0.048	0.055
Adjusted R^2	0.046	0.054

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors are used for the evaluation regression and standard errors for the allocation regressions are clustered at the participant level.