Norwegian School of Economics
Bergen, Fall 2020

Retail Investors' Preferences for Lottery Stocks

Are retail investors overinvesting in lottery-type stocks due to wealth constraints?

Jørgen H. Halvorsen Myhre and Thomas Gausemel Henriksen

Supervisor: José A. Albuquerque de Sousa

Master thesis, Economics and Business Administration

Major: Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

First and foremost, we would like to express our sincere gratitude to our supervisor, José A. Albuquerque de Sousa, for his thorough and constructive guidance throughout this process. His insights and feedback has truly been of great value. We also want to thank the Norwegian School of Economics for providing access to databases and literature that have been imperative for our research. Finally, we thank each other for a fruitful and enjoyable partnership in what has been a rewarding learning experience.

Norwegian School of Economics

Bergen, December 2020

Jørgen H. Halvorsen Myhre

Thomas Gausemel Henriksen

Abstract

We find that retail investors on aggregate are less attracted to stocks with highly skewed returns after an exogenous increase in stock tradability. Using data on the number of Robinhood investors owning U.S. stocks, we estimate the effect of the newly introduced Fractional Share Trading service to causally reduce retail investor holdings of such stocks by 30.78%. This finding is significant at the 10% level and is estimated using a Regression Discontinuity Design. Through a Difference-in-Difference analysis, we estimate a causal increase in the number of retail investors holding stocks that can be traded fractionally. However, when running a subgroup analysis of stocks with above median return skewness, retail investors' demand for such stocks is about 50% lower than for their low-skewed counterparts. Nevertheless, as both results from the DiD analysis are insignificant, they provide no conclusive answer. Lastly, using standard OLS, we observe that the significant positive relationship between skewness of stock returns and retail investment disappears once we control for stock tradability. Although these results are subject to uncertainty, all our findings suggest that wealth constraints can help explain why we observe an overrepresentation of retail investors in lottery-type stocks.

Keywords – Retail investors, Lottery stock preferences, Wealth constraints

Contents

List of Figures

List of Tables

1 Introduction

The purpose of this thesis is to analyse why retail investors are overly represented in lottery-type stocks. These are stocks that are (1) low-priced and (2) exhibit high return skewness. A portfolio of such stocks historically underperforms in the stock market by about 4% (Kumar, 2009) – a phenomenon which has been hypothesized to stem from retail investors' attraction to lottery-type payoffs. We hypothesize that budget constraints cause retail investors to be overexposed to stocks with high return skewness – causing their aggregate demand for lottery-type stocks to increase.

To test our hypothesis, we exploit the fact that the Robinhood trading platform only offers fractional share trading (FST) of stocks that report a market capitalization above $25 million and a price above $1 (Robinhood, 2020). This new service allows investors to purchase as little as one millionth of a stock, conditional on the stock satisfying the two criteria. Eligibility for this service allows for a Regression Discontinuity Design, in which we seek to establish whether there is a significant discontinuity in the number of Robinhood users holding stocks with highly skewed returns at the market capitalization threshold. Our choice to apply the market capitalization threshold is explained in greater detail in section 7.1. Using our RD design, we document a significantly negative causal treatment effect of 30.78% in retail investor holdings of stocks with the highest return skewness. However, this effect is only significant at the 10% level and when applying the broadest bandwidth.

We also include a Difference-in-Difference (DiD) design to test our hypothesis on retail investors' budget constraints. Through the DiD method, we compare the mean difference in the number of Robinhood users owning low- and high priced stocks before and after the introduction of FST. Our results yield a positive causal effect in the number of users holding high-priced stocks when they are traded fractionally, although the effect is non-significant. To extend on this result, we run a subgroup analysis to estimate the effect of FST on retail investor ownership of high-priced stocks with high return skewness. The treatment effect differs by about -50% for stocks with high return skewness compared to stocks with low return skewness. Thus, the causal increase in the number of users purchasing high-priced stocks after introducing fractional share trading is smaller for stocks with high return skewness. Although these results are in line with our hypothesis, they are insignificant, and we cannot give any conclusive answer using the DiD analyses.

Finally, we use OLS regression to test the relationship between stock return skewness and the number of Robinhood users holding stocks. Using eligibility for fractional share trading as our tradability proxy, the strongly significant and positive relationship between return skewness and number of retail investors holding stocks disappears once tradability is controlled for.

Although the results in both experimental designs are subject to considerable uncertainty, we observe the same trend across all three models: *retail investors shift towards stocks with less skewed returns once their capital limitations are reduced.* That is also the key takeaway of this thesis.

1.1 Motivation

So far, retail investors' excessive appetite for lottery-type payoffs have been thought to be a central cause of negative *lottery premiums* in the stock market. Stocks with highly skewed return features and low nominal prices are believed to be stock market equivalents of lottery tickets, and a portfolio of such stocks historically yields negative premiums (Bali et al., 2011). Several researchers have noted that retail investors are overly represented in these lottery-type stocks, which is why this group is thought to fuel the negative lottery premiums (Bali et al., 2011) (Han and Kumar, 2008). However, the view is split among researchers whether retail investors are drawn towards these stocks because they are wealth constrained towards low-priced stocks, or because they use low stock prices as a proxy for return skewness.

As there is a strong inverse relationship between nominal stock prices and return skewness, one cannot easily determine whether retail investors choose these stocks due to their perceived return skewness or their low price (Birru and Wang, 2016). We argue that retail investors are limited to a low-price investment universe that is inherently overweight in stocks with high return skewness. Their capital limitations will thus effectively alter their inclination towards stocks with high return skewness. We therefore seek to answer the following research question:

Do capital limitations cause retail investors to overinvest in stocks with highly skewed returns?

The Robinhood trading platform introduced Fractional Share Trading to their customers in late 2019, effectively removing the price barrier of stock investing. However, only stocks with a market capitalization equal to or above \$25 million and price above \$1 are eligible for being traded fractionally. If we hypothesize correctly, we expect to observe a decreased appetite for return skewness if investors are able to choose freely among stocks, independent of their nominal stock price.

1.2 Background

Based on the historical literature on gambling, it is likely that there exist investors that are willing to pay for lottery-type payoffs in the stock market. The prospect theory proposed by Kahneman and Tversky (1979) explains this observation by arguing that decision weights are assigned differently between gains and losses, which causes overweighting of the tails of a payoff distribution. This bias can help explain why one might see an overpricing of stocks with positively skewed returns, since skewed returns is a preference of which investors are willing to pay (Barberis and Huang, 2007).

To objectively observe stock market gambling, Kumar (2009) identifies stocks exhibiting lottery features as a proxy for observing gambling-motivated behavior in the stock market. Based on insights gathered from state lottery tickets and studies, he defines lottery stocks as stocks exhibiting (1) low price, (2) high idiosyncratic volatility and (3) high idiosyncratic skewness in their returns. However, research by Bali et al. (2011) find that, when controlling for past maximum returns, idiosyncratic volatility turns out insignificant. This finding implies that idiosyncratic volatility merely proxies for skewness. Accordingly, we do not consider idiosyncratic volatility as a lottery stock characteristic moving forward. Low-priced stocks with highly skewed returns are therefore the lottery stocks we will be researching in this paper.

In their early-stage empirical results, Downs and Wen (2001) document the existence of *"lottery premiums"* – the willing sacrifice in average returns that investors pay for a remote chance to earn an abnormally positive return. They also find overpricing to be prevalent in low-priced stocks, and that the negative premiums associated with these stocks tend to increase as the nominal stock price decrease. Kumar (2009) later extended the definition of lottery stocks to also include high return skewness and high volatility. He further documents a -4% annual stock premium for such stocks in the American stock market.

2 Literature review

The existing literature collectively hypothesize that retail investors might be a driving force behind lottery premiums[1]. Several studies find correlations between lottery stocks and retail investors. For example, Kumar (2009) finds that preferences for lottery stocks are dominant among individual investors and that institutional investors show a preference against lottery-type stocks. Bali et al. (2011) complement these findings by noting that retail investors exhibit higher preferences for lottery-like stocks. However, as lottery-like stocks are defined as having low prices and high return skewness, the fact that these characteristics are interrelated complicates the assumption on retail investment preferences for such assets. In other words, to what extent retail investors seek low prices and skewed returns as a joint combination is not straightforward.

One reason for retail investors' overweight in lottery-type stocks might be their capital limitations, and not an excessive preference for lottery-type payoffs. Proponents of the *Marketability Hypothesis* argue that retail investors are limited to low-priced stocks, providing an incentive for management to lower their stock prices through stock splits as a marketability scheme[2]. This hypothesis is based on a diversification argument, as share prices cannot rise too high before retail investors need a significant amount of capital to diversify a portfolio of high-priced stocks (Baker and Gallagher, 1980). If true, there might be a case for reverse causality in the argument of retail investors seeking lottery-type payoffs. Because stocks with low prices are found to have inherently more skewed returns than high-priced stocks, this relationship suggests that retail investors are effectively limited to an investment universe that is more lottery-like than others (Benartzi et al., 2009) (Birru and Wang, 2016).

On the other hand, retail investors might concentrate in low-priced stocks exactly because these stocks have higher return skewness than high-priced stocks. Research on nominal prices find that investors tend to overestimate the return skewness of low-priced stocks, potentially enhancing any preferences towards such stocks, as this characteristic will proxy for skewness. This finding is rooted in a biological bias, in which investors often view a low-priced stock as having more upside potential than a high-priced stock, because

[1] See for example Han and Kumar (2008), Kumar (2009) or Bali et al. (2011).

[2] See Schultz (2000), Easley et al. (2001) and Dahr et al. (2004).

the former is closer to zero and farther from infinity (Birru and Wang, 2016).

However, the Achilles' heel of the lottery stock portfolio set forward by recent literature, is the way in which retail investors are expected to identify the lottery-stocks to pursue, given that high return skewness in fact is their preference. This view implies that investors need to somehow estimate and compare the skewness of stock returns in the market. Kumar (2009) argues that *"certain salient stock characteristics"*, such as industry or media coverage, might be one way to do so. However, he recognizes that extrapolating previous maximum returns into the future, and then comparing nominal stock prices is the most likely way in which retail investors identify and compare skewness.

While there is reason to believe that retail investors follow the maximum return strategy to a certain extent, research on nominal stock prices suggests that retail investors tend to categorize stocks based on price, implying that stock prices are important to retail investors (Green and Hwang, 2009). Support for this view is found in several studies conducted on stock splits[3]. Dyl and Elliott (2006) complement these findings by adding that retail investors tend to hold lower priced stocks than institutions. (Benartzi et al., 2009). If retail investors in fact categorize stocks based on their price, the previously mentioned inverse relationship between skewness and stock prices will ensure that their investment categories will differ in terms of return skewness.

Based on the reviewed literature, it is difficult to answer whether preferences for lottery-like stocks stem from preferences for skewed returns or a constraint towards low prices. While investors certainly can have extraordinarily high preferences for skewness while also having budget constraints, budget-constrained retail investors are nonetheless biased towards lottery-like stocks due to the inverse relationship between nominal prices and return skewness. Thus, retail investors on an aggregate level might be more inclined to purchase stocks with highly skewed returns as a consequence of their budget constraints.

In sum, whether retail investors predominantly pursue lottery stocks due to their highly skewed features, or due to their budget constraining them to low-priced stocks, remains an unanswered question which this thesis will try to answer.

[3] See Schultz (2000), Easley et al. (2001) and Dahr et al. (2004).

3 Hypotheses

To follow up on the views presented in the literature review, there is a mixture of hypotheses why retail investors concentrate in lottery stocks. Whereas low nominal prices could be a lottery preference due to the illusion of skewed returns, it might also be a preference due to the fact that it allows for diversification for investors with limited capital. We therefore have two main hypotheses that might explain retail investors' preferences for lottery-type stocks.

Due to wealth constraints, retail investors' portfolios are effectively limited to a low-priced universe of stocks – potentially causing them to overinvest in lottery-type stocks, as these are low-priced by definition. Therefore, we expect retail investors to show less preference for stocks with highly skewed returns if given access to a wider investment universe through the introduction of fractional share trading. We will call this hypothesis the *Wealth Constraint Hypothesis* moving forward.

A second hypothesis however, is that retail investors in fact value skewness more than price. If that is the case, retail investors should exhibit significantly increased attraction towards stocks with highly skewed returns, given that they are able to trade fractionally. We will call this hypothesis the *Lottery Preference Hypothesis.*

4 Methodology

We use four different methodologies in attempting to answer our research question. First, inspired by Kumar (2009), we investigate the existence of lottery stock premiums in the American stock market during the last two decades, as the existence of such premiums serves as the essence of our thesis. To do so, we use the two-factor model of Harvey and Siddique (2000) to measure the characteristics necessary to identify lottery-type stocks, i.e. idiosyncratic volatility and idiosyncratic skewness. Following our argumentation in section 1.2, we use a redefined definition of lottery stocks, sorting our lottery portfolio solely based on price and idiosyncratic skewness. Using the sorted lottery portfolio, we run cross-sectional regressions by method of Fama and MacBeth (1973) to estimate the lottery premiums.

Included in our robustness analysis however, we also regress a lottery-portfolio sorted on the three traditional characteristics (Kumar, 2009), for means of comparison and to ensure the validity of our results. In doing so, we make use of Fama & French's three-factor model (Fama and French, 1993). As we also apply their asset pricing model to estimate the three risk betas to use as controls in our regression analyses, the three-factor model is briefly introduced in section 4.1.

Second, we exploit the fact that Robinhood does not offer Fractional Share Trading for stocks with market capitalization below $25 million. This threshold allows us to investigate retail investors' preferences between skewed payoffs and nominal stock prices, by comparing their stock holdings at the market capitalization threshold. Using data on the number of Robinhood investors holding U.S. stocks, we apply a Regression Discontinuity Design to determine whether there is a discontinuity in the number of retail investors holding stocks with highly skewed returns at the threshold.

Third, we use the Difference-in-Difference framework to analyse the mean number of retail investors holding stocks before and after the introduction of FST. Thus, we compare the difference in the average number of Robinhood investors holding low- and high-priced stocks before and after the introduction of FST. To test whether the FST implementation had a different effect on stocks with high return skewness, we extend this model to include an interaction term for stocks with above median skewness in their returns.

Lastly, we use OLS to estimate the correlation between retail investment and stock return skewness. If our hypothesis is correct, we would expect to see a decreased preference for stocks with skewed returns when we control for stock tradability.

4.1 Fama & French's Three-Factor model

Fama & French's 3-factor asset pricing model is an extension of the CAPM, including risk proxies for size and value in addition to the market factor. By regressing stock market returns on common risk factors for assets, this asset pricing model is empirically able to explain up to 95% of variation in asset returns (Fama and French, 1993). In the following, we will briefly explain the variables used and the method for which one can validate these findings.

The three-factor model is named after three risk variables that is thought to explain asset returns: market risk, size risk and value risk. The extension of this model from the CAPM comes from adding the two latter risk factors; company size- and value. These factors are included as a response to the empirical observation that small-cap stocks and value-stocks outperform the rest of the market (Fama and French, 1993).

The market factor is computed as the daily excess return on the market of all common stocks traded on NYSE, NASDAQ and AMEX, weighted by each stock's respective market values of June each year. As firms across different industries are subject to different accounting regulations and deadlines, they use market values on common reference dates each year to ensure comparability (Fama and French, 1993).

To obtain these estimates, Fama & French sort each stock by size and book-to-market ratio on the last trading day each June, using their accounting data for December of the previous fiscal year. These portfolios are then held until June of year *t+1*, at which point the sorting repeats. Stocks with market capitalization below the median are assigned as small-cap *(S)*, the rest as large-cap *(B)*. They use the .3 - and .7-quantiles to sort stocks on book-to-market ratio. The top 30% are classified as firms with high book-to-market value *(H)*, the middle 40% as medium *(M)* and the bottom 30% as low *(L)*. Lastly, using these intersections of size and value, they form six value-weighted portfolios. *S/L*, *S/M*, *S/H*, *B/L*, *B/M* and *B/H*. To derive the stock market return factors, they compute the value-weighted returns of the two zero-net-investment portfolios:

$$R_{SMB} = \frac{1}{3}(R_{S}/_{L} + R_{S}/_{M} + R_{S}/_{H}) - \frac{1}{3}(R_{B}/_{L} + R_{B}/_{M} + R_{B}/_{H}) \tag{4.1}$$

$$R_{HML} = \frac{1}{2}(R_{S}/_{H} + R_{B}/_{H}) - \frac{1}{2}(R_{S}/_{L} + R_{B}/_{L}) \tag{4.2}$$

SMB is the excess return on a zero-net-investment portfolio of small stocks minus a portfolio of large-cap stocks, and *HML* is the excess return on a zero-net-investment portfolio of high market-to-book ratio minus a portfolio of low market-to-book ratio. Thus, these factors are added to the CAPM asset pricing model as *SMB* (Small minus Big) and *HML* (High minus Low) (Fama and French, 1993). The final three-factor asset pricing model can therefore be expressed as:

$$R_f - r_i = \alpha_i + \beta_i(R_m - R_f) + s_i SMB + h_i HML + \epsilon_i \tag{4.3}$$

4.1.1 Lottery stock sorting

In order to construct the lottery-type portfolio, we identify stocks that (1) have low price and (2) highly skewed returns. From the literature, low price is defined as below median and high skewness is defined as above median. To measure the idiosyncratic skewness, we adopt the established methodology of Harvey and Siddique (2000), and estimate the two-factor model of the excess and squared market return measured over the past six months:

$$r_{i,d} = \alpha_i + \beta_i MKT_d + \gamma_i MKT_d^2 + \epsilon_{i,d} \tag{4.4}$$

First, we compute *IVOL*, which is derived as the standard deviation of the residuals obtained by fitting the two-factor model:

$$IVOL_{i,t} = \sqrt{\frac{1}{N_t} \sum_{d \in S_t} = \epsilon_{i,d}^2} \tag{4.5}$$

Next, we can derive the idiosyncratic skewness, *ISKEW*:

$$ISKEW_{i,t} = \frac{1}{N_t} \frac{\sum_{d \in S_t} \epsilon_{i,d}^3}{IVOL_{i,t}^3} \tag{4.6}$$

where Nt denotes the number of trading days and St is number of days in month t

After computing the idiosyncratic skewness of each stock observation, stocks are sorted into portfolios of low *(L)* or high *(H)* price and low *(L)* or high *(H)* idiosyncratic skewness. The *Lottery-Type* portfolio is defined as *(L/H)*, and the *Non-Lottery* portfolio is defined as *(H/L)*. The remaining portfolios, *(H/H)* and *(L/L)*, are defined as *Other*. By adding these to the three-factor model of Fama & French, we can achieve estimates for any premium on these portfolios. The final model of our lottery-type sorting is thus defined as:

$$R_f - r_i = \alpha_i + \beta_i(R_m - R_f) + s_i SMB + h_i HML + l_i \mathit{Lottery\text{-}Type} + \tag{4.7}$$

$$nl_i \mathit{Non\text{-}Lottery\text{-}Type} + o_i \mathit{Other} + \epsilon_i$$

5 Data

From Compustat we obtain daily stock data from 2000 to 2020. For our analysis on lottery stock premiums, we only use U.S.-based common stocks traded on NYSE, Nasdaq and AMEX (Wharton Research Data Services, 2020). We also retrieve the data necessary to compute the book equity of each stock. The detailed process of screening, cleaning and matching these data can be found in section A1 of the appendix, while all variables in our final data set are explained in section A2.

In order to investigate retail investors' preferences, we also download popularity data from Robinhood, reporting the daily number of users holding each stock (Robintrack, 2020). Using this data in conjunction with the daily stock data from Compustat, we are able to observe the day-by-day developments in ownership of all stocks in our sample, thus allowing us to perform aggregate analyses on retail investor behaviour. However, as the Robinhood platform is fairly new, this data is only available from 2018 to present.

5.1 Limitations

The main analysis in this paper is limited to only the first six months of 2020, as the Fractional Share Trading service was introduced in late December 2019. In combination with lottery-type stocks by definition being rare, we are therefore left with a relatively small data sample. With limited sample size, interpretation and inference of our estimates and findings from the analyses thus call for conservatism.

Moreover, eligibility for fractional share trading is conditional on rules put forth by Robinhood – namely that each stock must trade above \$1 and have equal to or higher than \$25 million in market capitalization. As we are unable to directly observe whether all stocks that satisfy these two rules are being traded fractionally, our analyses might be corrupted by non-compliance.

6 Replication of lottery stock premiums

As the scope of our paper is to investigate retail investors' preferences for lottery-type stocks, documenting the existence of lottery stock premiums is imperative. This part therefore replicates the efforts of previous research related to estimating such premiums. In addition, this process is also crucial to identify lottery-type stocks and to construct the lottery portfolio needed for our main analysis. In the following, we report the results of our estimation of such premiums in the American stock market. For additional analyses on lottery portfolio performance, time-series effects and robustness analysis, we refer to sections B1 - B5 of the appendix.

Since Kumar (2009) have previously identified negative lottery premiums in the American stock market, and due to Vokatá (2012) reporting the existence of lottery stock premiums on several of the big exchanges across Europe, we expect similar findings. In accordance with the lottery stock sorting described in 4.1.1, we sort our sample in three different portfolios based on stock price and idiosyncratic skewness. With basis in the lottery portfolio, we conduct cross-sectional regressions to investigate the lottery stock premiums across our sample period, using the method of Fama and MacBeth (1973).

6.1 Fama-MacBeth Two-Step Regression

The Fama and MacBeth (1973) regression is a two-step approach allowing us to observe risk premiums over time, and deserves closer explanation. First, each individual excess stock return is regressed against the time-series of the included explanatory variables, in order to determine its exposure to each factor. Second, the cross-section of excess returns is regressed once more against the factor exposures at each time t, yielding a time-series of risk premia coefficients for each factor. The key insight of the Fama-MacBeth-approach is to then average these coefficients, one for each factor, to obtain the premium expected for a unit exposure to each risk factor over time. To achieve this premium we use the following model specification:

$$R_{i,t} = \alpha_{i,t} + x_{i,t} LotteryDummy + \beta_{i,t} MKT + s_{i,t} SMB + h_{i,t} HML + \tag{6.1}$$

$$x_{i,t}Log\_mktcap + x_{i,t}bm + x_{i,t}LaggedR + \epsilon_{i,t}$$

In order to isolate the effect of lottery-type stocks on value-weighted excess returns, we specify a dummy variable *(LotteryDummy)*, taking the value of 1 if a stock has low price and returns with high idiosyncratic skewness. Hence, it is in particular the coefficient sign and significance level of the *LotteryDummy* that are of interest to us. Value-weighted excess stock return in month *t+1* is the dependent variable, while monthly risk factors (*mkt*, *smb* and *hml*), the log of market size, book-to-market ratio and lagged 6-month return are included in the model as controls.

Contrary to traditional panel data regressions – in which the common approach is to use robust, clustered standard errors (HAC) to correct for heteroskedasticity and autocorrelation – the procedure applied in the Fama-MacBeth-regression is considered a special case. As mentioned, we estimate the regression separately for each month, before testing hypotheses about the coefficient of interest by the t-statistic of the resulting monthly coefficient estimates. Following the paper of Ibragimov and Muller (2010), the Fama-MacBeth method yields valid inference as long as the monthly coefficient estimators are approximately normal and independent, even on short panels that is heterogenous over time.

6.2 Results

Table 6.1 reports evidence of negative and significant lottery premiums on our lottery-type portfolio – which constitutes of stocks with price below median *(L)* and above median return skewness *(H)*. The monthly lottery premium, represented by the estimated coefficient of the *LotteryDummy*, equals -0.501, which adds up to -6,01% on an annual basis. This finding is thus consistent with the previous research of Kumar (2009), who documented similar annual premiums on the American stock market of at least -4%.

Table 6.1: Premiums on lottery-type portfolio

This table reports the Fama and MacBeth (1973) monthly cross-sectional regression estimates for the value-weighted excess return on the lottery-type portfolio. The main variable of interest is the lottery dummy, assigning the value of 1 to all stocks that form our lottery portfolio. Contemporaneous factor betas for the Fama-French risk factors (mkt, smb, hml), size (log of market value), book-to-market and lagged 6-month return are included as controls. t-statistics are reported in parentheses.

	Dependent variable:
	Value-weighted return on lottery-portfolio
LotteryDummy	-0.501^{***} (-5.386)
mkt	0.937^{***} (5.008)
smb	-0.309 (-0.874)
hml	0.977^{***} (7.584)
Log_mktcap	2.133^{***} (16.333)
bm	0.883^{***} (11.852)
LaggedR	-0.011^{**} (-2.500)
Constant	13.465^{***} (6.207)
Model	Fama-MacBeth
Observations	671,039
R^2	0.230
Note:	*p<0.1; **p<0.05; ***p<0.01

The market factor coefficient, *mkt*, is significant at the 1% level, and almost equal to 1, as we would expect since our stock sample is a reasonable representation of the market. Further, in our stock sample, high-value stocks receive positive and significant return compensations, while the effect of small-cap stocks are negative, albeit not significant.

In line with the predictions of Fama and French (1993) however, we would expect small-cap stocks to outperform large-cap stocks - as they have done historically. One plausible explanation why we observe the opposite relationship could be rooted in the evidence of Switzer (2010), in which he documents that small-cap stocks tend to lag large-cap stocks in periods leading up to business cycle peaks. Another possibility is that it could be ascribed to the scope of the sample period included in our paper. As analyst Erik Norland highlights, large-cap stocks (proxied by S&P 500) - mainly due to sustained periods of economic expansion - have outperformed small-cap stocks (proxied by Russell 2000) since the end of 2013 until 2020 (Norland, 2020). In light of this observation, it therefore makes sense that we observe the estimated coefficient for *smb* to be significantly positive between 1999-2010, while it shifts to a negative sign when estimating it across the full sample period. See table B5.1 in section B5 of the appendix for the complete robustness analysis of our stock sample.

Stocks with past high returns (measured over the previous 6 months) however, earn significantly lower returns. It seems reasonable to assume that stocks that exhibit high returns will be subject to increased scrutiny and attention among retail investors. Consequently, high returns could increase the possibility of overpricing in the next period – which ultimately implies lower returns. The coefficient of *LaggedR* therefore have the expected negative sign. We also note that *Log_mktcap* and *bm*, proxying for size and value respectively, are both positive and significant. Finally, we report an R^2 of 23%, which means about 1/4 of the variation in monthly value-weighted returns is explained in the model.

6.3 Descriptive statistics

To better understand the nature of the stocks in our lottery portfolio, and their inherent features, we provide summary statistics of a broad range of firm-characteristics in Table 6.2 by method of Kumar (2009). For means of comparison, the non-lottery-type and other portfolio are also included.

Table 6.2: Descriptive statistics of lottery-type portfolio

This table reports monthly mean characteristics of lottery-type stocks measured over the sample period from 2000 to 2020. Descriptive statistics of stocks in the non-lottery-type and other portfolio are also reported for means of comparison. All characteristics represent averages measured over the sample period.

	Lottery-Type	Non-Lottery-Type	Other
Value-weighted return (in %)	-1.50	1.03	-0.01
Number of stocks	511	495	1,733
Market share (in %)	2.33	39.19	58.48
Idiosyncratic volatility	3.84	1.23	2.17
Idiosyncratic skewness	1.13	-0.69	0.04
Stock price	9.06	56.15	32.97
Market beta	1.03	0.72	0.88
SMB beta	0.97	0.20	0.53
HML beta	0.27	0.40	0.29
Firm size	768	13,677	5,843
Book-to-Market	1.10	0.57	0.79
Past MAX return (in %)	11.24	2.80	4.89
Past 12-month return (in %)	-18.17	4.54	-2.68

On average, we observe that the lottery-type portfolio only account for just above 2% of the total market, against almost 40% for the non-lottery-type portfolio. This observation indicates that the lottery-type portfolio mainly consists of small-cap stocks. We also draw a similar conclusion when comparing the averages of firm size across the three portfolios.

Moreover, while lottery-type stocks on average earn considerably lower past 12-month returns, they however exhibit considerably higher average *MAX*, measured as the maximum return observed over the previous month *(t-1)*. In conjunction, these two observed features suggest that the stocks in the lottery-type portfolio behaves in accordance with findings of previous research – lottery stocks perform worse on average, but at the same time exhibit a more pronounced potential of yielding extreme positive returns.

6.4 Univariate analysis

Motivated by our preliminary findings, we further investigate our hypotheses by looking for trends and patterns in our stock sample. To do so, we conduct two univariate analyses.

We sort stocks into price deciles, and report the mean statistics of stock characteristics for each decile. We also report summary statistics for Robinhood investors' stock holdings - from the least to the most popular decile.

By sorting stocks based on their price, we are able to observe whether there are inherent differences in stock characteristics between low-priced and high-priced stocks. In table 6.3, we provide summary statistics for the price deciles.

Table 6.3: Univariate analysis on price

This table reports mean characteristics of our stock sample, sorted in deciles by stock price. The deciles are reported from lowest to highest.

	Price	IVOL	ISKEW	Lottery	MktCap	BM	UsersHolding	N
LowDecile	1.143	7.451	0.733	0.716	23.584	1.728	144,445	109
D2	1.465	5.744	0.296	0.580	25.029	1.670	142,135	100
D3	1.896	5.647	0.497	0.625	25.865	1.753	145,964	104
D4	2.350	5.559	0.412	0.594	24.057	2.361	77,390	101
D5	2.823	4.886	0.323	0.624	27.047	2.224	50,314	101
D6	3.436	5.387	0.579	0.584	23.946	1.776	45,410	101
D7	4.260	5.046	0.245	0.594	30.718	1.256	56,677	101
D8	6.114	4.363	0.215	0.505	30.421	1.136	62,787	105
D9	9.772	3.959	0.094	0.446	31.032	1.284	21,486	101
HighDecile	21.067	3.501	0.074	0.485	35.002	1.733	10,306	101

Observing the column for *UsersHolding*, which reports the total number of Robinhood investors in each decile, there is a clear trend of attraction towards the low-priced stocks. We also observe that the average skewness in returns of low-priced *(LowDecile)* stocks is higher than for high-priced stocks *(HighDecile)*. By conducting a Welch two sample t-test of equality in group means, we reject the null hypothesis of equality, with a reported t-statistic of 3.629. Thus, this difference is significant in both an economical and a statistical sense. Moreover, this pattern is consistent with the findings of Birru and Wang (2016), who find that low-priced stocks are inherently more skewed in their returns than others. Hence, these observations are supportive of the *Wealth Constraint Hypothesis*. The relationship between price and market capitalization also moves in the opposite direction, which implies that high-cap stocks tend to have higher stock prices. *Lottery* is a lottery stock indicator, and increases with high values of *ISKEW* and low values of *Price*.

To better understand the holdings of Robinhood investors, we also group the stocks into deciles based on the number of investors holding each stock. In table 6.4 we report the mean statistics for each decile.

Table 6.4: Univariate analysis on Robinhood investors' stock holdings

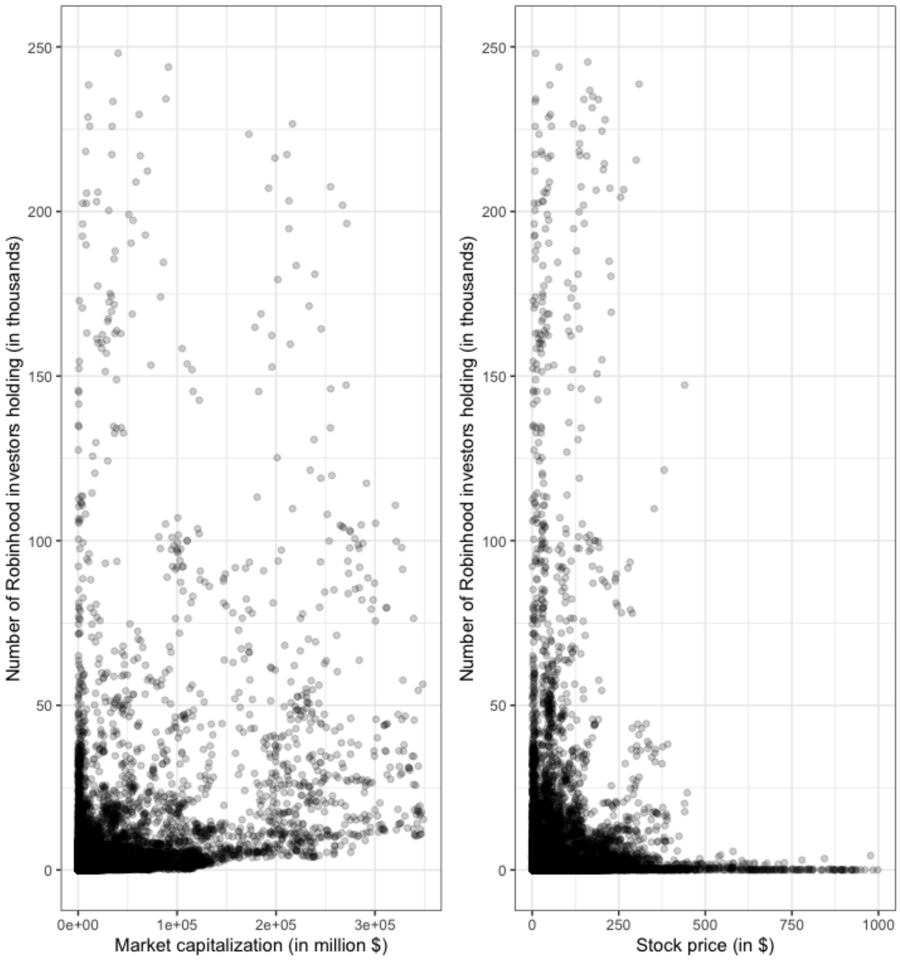
This table reports mean characteristics of our stock sample, sorted in deciles based on Robinhood investors' holdings. The deciles are reported from lowest to highest.

	UsersHolding	IVOL	ISKEW	Price	Lottery	MktCap	BM	N
LowDecile	1,941	3.347	0.017	16.686	0.453	31.637	1.806	106
D2	4,940	3.913	0.290	9.640	0.524	30.424	1.334	103
D3	10,031	4.552	0.156	5.255	0.485	27.195	1.152	101
D4	16,918	5.413	0.335	4.047	0.618	26.392	1.660	102
D5	27,004	5.536	0.551	3.561	0.604	26.247	2.059	106
D6	34,077	5.298	0.381	3.253	0.600	31.668	2.426	100
D7	47,363	6.100	0.583	3.351	0.616	21.955	2.164	99
D8	75,945	5.560	0.347	2.511	0.613	25.367	1.458	106
D9	115,572	5.923	0.349	2.540	0.626	28.180	1.517	99
HighDecile	423,123	6.159	0.500	2.622	0.627	27.280	1.338	102

Inspecting table 6.4, there are several trends in the summary statistics worth noting. First, there seems to be a negative relationship between the number of users holding stocks and stock prices – consistent with the findings in table 6.3. Second, we observe a positive relationship between the number of users holding stocks and the idiosyncratic skewness, as the mean of *ISKEW* is higher for stocks with the highest number of users holding, *HighDecile.* We test and find the difference in means of *ISKEW* between the *LowDecile* and *HighDecile* to be significant, with a t-statistic of 2.642. This observed pattern thus seem to suggest that stocks have significantly higher return skewness when their retail investor ownership is high. This observation provides support to our proposed *Lottery Preference Hypothesis.* Interestingly, we also observe that amount of lottery-type stocks is lowest in the *LowDecile.*

Expanding further on our univariate analysis, we also plot the number of Robinhood investors holding against both market capitalization and stock price. From figure 6.1, we learn that retail investors on the Robinhood-platform concentrate in stocks that have (1) low market capitalization and (2) low stock price. In particular, the concentration of retail ownership is most pronounced when the stock price is arbitrarily close to zero. These observations are thus consistent the views of Benartzi et al. (2009) who argue that retail investors are limited from purchasing high-priced stocks due to wealth constraints. In conjunction with the observed pattern from table 6.3 - in which idiosyncratic skewness is significantly higher among low-priced stocks - the observed behavior of retail investor concentration in low-priced stocks is also consistent with the findings of Kumar (2009), who document that retail investors are overly represented in lottery-type stocks.

Figure 6.1: Robinhood investors' stock holdings, market capitalization and stock price



This figure plots the observed number of Robinhood investors holding each stock against the stocks' market capitalization and stock price, respectively.

By conducting univariate analyses on both price and the number of Robinhood investors holding each stock, we have identified some notable trends in our stock sample. Sorting on stock price, we reveal a negative trend between price and idiosyncratic skewness – suggestive of low-priced stocks having inherently more skewed return features than high-priced stocks on average. Moreover, when we sort on the number of users holding, we learn that there seem to exist a negative relationship between the number of users holding and stock price. Observing the idiosyncratic skewness across the deciles, we also observe that idiosyncratic skewness tends to be higher for the stocks with higher retail ownership.

7 Lottery stock preferences and stock tradability

As established in the literature review, the consensus seem to be that retail investors are the dominant cause of lottery stock premiums (Kumar, 2009) (Bali et al., 2011). Having documented the existence of such negative lottery stock premiums, we are subsequently interested in determining what really causes retail investors to gravitate towards these stocks. In this part, we therefore revisit our research question from the introduction – seeking answers to whether it is the low price, or the inherent skewness in returns of lottery stocks that attracts retail investors.

Through retail broker Robinhood's introduction of the Fractional Share Trading service, we are able to observe retail investor behaviour without a nominal price barrier – as it enables retail investors to invest fractionally in stocks, widening their investment universe to include stocks that were previously too expensive. As we have identified the lottery stocks in the American market, we are able to compare these stocks to the ones Robinhood's investors are holding. Since FST is only offered for stocks reporting a market capitalization of at least $25 million and price above $1, we exploit these thresholds to test whether preferences for lottery-type stocks among retail investors persist after we control for stock tradability.

7.1 Regression Discontinuity Design

To test the effects of fractional share trading on retail investors' preferences, we apply the method of Regression Discontinuity Design (Lee and Lemieux, 2010). Accordingly, we exploit that eligibility for the service is conditional on passing the threshold of $25 million and a trading price above $1. This analysis will however focus on the threshold of $25 million, as it is reasonable to assume that retail investors' capital constraint threshold lies above $1. The purpose of this research design is therefore to analyse whether there is a significant discontinuity in the average number of retail investors holding stocks with skewed returns at the threshold of $25 million. Stocks above this threshold can be bought fractionally, whereas stocks below cannot.

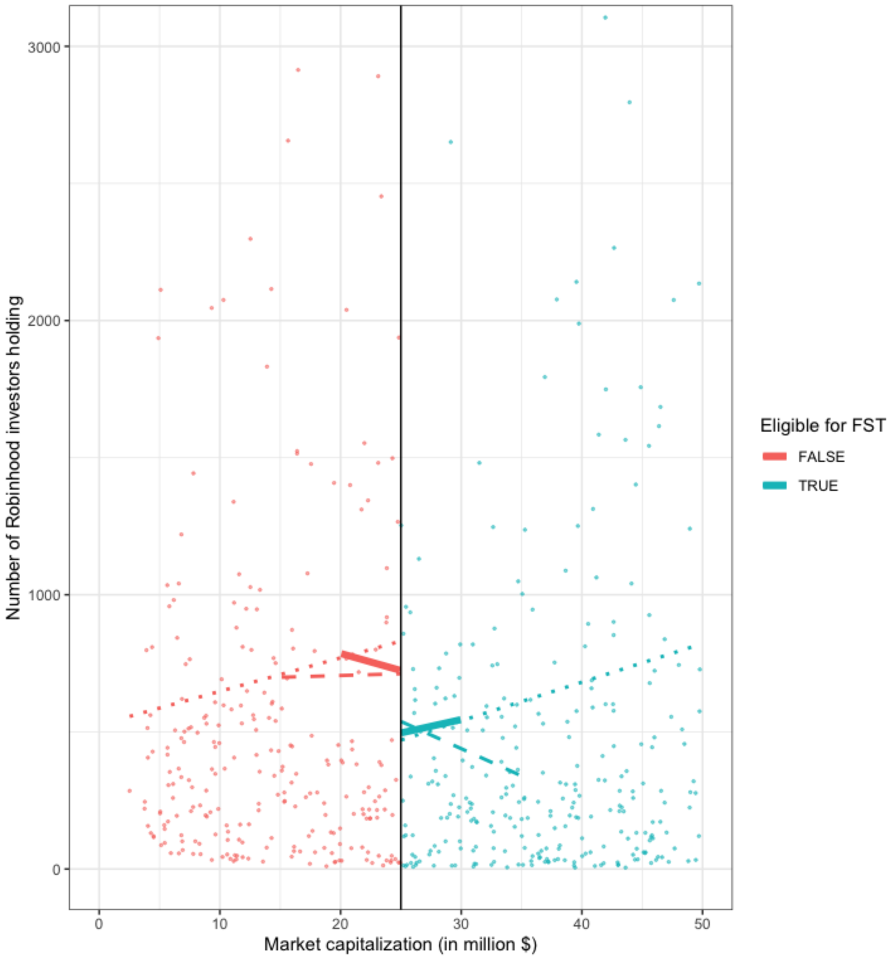
As fractional share trading removes the price barrier for retail investors, we exclude price as a lottery stock characteristic. Thus, our lottery portfolio is reconstructed by only including stocks with above median idiosyncratic skewness. Furthermore, we exclude all stocks with price below \$1 as they are ineligible by definition. Stocks with market capitalization above \$50 million are also excluded from the analysis, as we consider these observations as too different from the ones near the cut-point. The considerations related to the comparability of eligible and ineligible stocks above and below the threshold will be discussed in greater detail in section 7.4 when we assess the internal validity of our RD design. Lastly, as the FST-service was implemented at December 16th, 2019, we use data on Robinhood investors' stock holdings from this date and forward to test the effect.

7.2 Graphical presentation

In figure 7.1, we plot the relationship between observed values of market capitalization and the number of Robinhood investors holding each stock with return skewness above median. We apply a bin size of 2 and market capitalization bandwidths of \$5, \$10 and \$25 million. These considerations are discussed further in section 7.4 and depicted in C2.

The graph in figure 7.1 shows a clear discontinuity near the cut-point for all bandwidths. However, the relationship between users holding and market capitalization is not consistent across all bandwidths. This inconsistency is a great concern for our model, questioning the trustworthiness of our simple functional form. However, as the number of observations is reduced significantly when changing the bandwidth to 10 or 5, one should be aware of the uncertainty in these estimates. Nevertheless, the discontinuity in the cut-off remains consistent across all bandwidths.

Figure 7.1: Number of Robinhood investors holding stocks with above median return skewness above and below the threshold of eligibility for fractional share trading



This figure shows the observed number of Robinhood investors holding stocks with above median return skewness, above and below the eligibility threshold of $25M, under three different bandwidth specifications. The dotted line represents a bandwidth of 25, the dashed line a bandwidth of 10 and the bold line a bandwidth of 1. The vertical line marks the cut-point for fractional share trading eligibility.

7.3 Retail investors' preferences and stock tradability

Proxying for retail investors' preferences, and included as our dependent variable, we use the number of Robinhood users holding each individual stock with above median return skewness. As the distribution of *Users holding* is positively skewed, it is log-transformed to mitigate the effect of potential outliers. This transformation changes the interpretation of the regression coefficients, in which a unit increase in our explanatory variables yield the approximate percentage change in *Users holding*. Further, we specify a dummy variable, *EligibleDummy*, taking the value of 1 if market capitalization is above $25 million, and zero otherwise. Having excluded all stocks with prices below $1, market capitalization is our sole running variable. Lastly, *mktcap_ centered* represents the distance in observed market capitalization from the threshold of $25 million, while *Price* and *MAX* (past maximum returns) are included as control variables. The extensive step-by-step approach of specifying our regression model is explained in section C1 of the appendix.

We run four separate linear regressions – each with different bandwidths. The results are summarized in table 7.1 below.

Table 7.1: Regression Discontinuity on lottery stock preferences

This table reports the multiple linear regression estimates for the logarithm of the number of users holding, proxying for retail investors' preferences, on market capitalization (the running variable), an eligibility dummy indicating stock tradability of fractional share trading, stock price, past maximum return measured over the previous month and past month's observed idiosyncratic skewness. Models (1) to (4) are specified using different bandwidths. Market capitalization is centred to indicate the distance of each observation from the cut-point. T-statistics are reported in parentheses.

	Dependent variable:			
	log(Users holding)			
	(1)	(2)	(3)	(4)
EligibleDummy	-0.368^{*}	-0.104	-0.037	-0.389
	(-1.761)	(-0.349)	(-0.089)	(-0.391)
mktcap_centered	0.016^{**}	-0.0003	-0.012	0.431
	(2.068)	(-0.011)	(-0.159)	(0.493)
Price	-0.104^{***}	-0.125^{***}	-0.173^{***}	-0.159^{***}
	(-12.725)	(-9.297)	(-8.054)	(-4.113)
MAX	0.013^{***}	0.016^{***}	0.018^{**}	0.022
	(3.654)	(3.019)	(2.594)	(1.208)
lagged_iskew	0.209^{***}	0.109	-0.014	-0.158
	(3.444)	(1.200)	(-0.126)	(-0.667)
Constant	5.886^{***}	5.827^{***}	6.188^{***}	6.299^{***}
	(40.250)	(26.951)	(21.320)	(8.185)
Model	OLS	OLS	OLS	OLS
Bandwidth	25	10	5	1
Observations	552	230	125	28
Adjusted R^2	0.272	0.316	0.379	0.383
Note:	*p<0.1; **p<0.05; ***p<0.01			

Despite being negative across all four models, the causal treatment effect on *log(Users holding)*, measured by the coefficient of the *EligibleDummy*, is only significant when we apply the widest bandwidth – questioning the validity of the inference we can derive from this analysis. While we cannot decisively conclude that the treatment effect is significant, all models indicate a negative relationship between *log(Users holding)* and the *EligibleDummy*.

Accounting for the number of observations reported across the four models, a great concern of our RD design is the small sample size, which equals 552 observations at most. This concern is evident when we narrow our bandwidth from 25, regressing our models on only 230, 125 and 28 observations respectively. In light of the findings of Jacob et al. (2012), which is commented in section 7.4.7, basing our RD design on a considerable sample size is crucial for obtaining precise estimates. On the other hand, the narrower the bandwidth we apply, the more representative the causal effect will be of the stocks near the threshold. Faced with a weigh-off between sample size and choice of bandwidth, our findings require caution when trying to conclude on any causal relationship between preferences for return skewness and stock tradability. However, by following the recommendations of Jacob et al. (2012), we give priority to sample size, placing more trust in the estimates derived from model (1). Thus, interpreting the estimated coefficient of *EligibleDummy*, we find evidence suggesting that retail investors' exhibit a preference against lottery-type stocks when we control for stock tradability. Since we have log-transformed our dependent variable, the reduction in lottery-stock preferences on aggregate equals -30.79% ($\approx (e^{(-0.368)} - 1) * 100$).

The coefficient for *mktcap_ centered* represents the percentage change in number of users holding, per \$1M unit increase in market capitalization above the threshold. For example, in model (1) we expect the number of users holding to increase by 1.61% ($\approx (e^{(0.016)} - 1) * 100$), for every \$1M unit increase in market capitalization above \$25M. The included controls, *Price* and *MAX*, are both significant and show the expected signs across all four models. Higher prices are associated with a significantly lower number of users holding, while past extreme returns show the opposite relationship. In attempting to ascribe retail investors' preferences for lottery-type stocks to either low prices or highly skewed return features, we choose to concentrate our analysis around the two lottery characteristics. Accordingly, there will be certain factors, such as interest rates, taxation policies, analyst coverage and so on, which also affect investor preferences. These factors are unaccounted for in our model and thus constitute the error term.

7.4 Considerations and internal validity

7.4.1 Optimal bin width

To avoid any potential bias from specifying a too narrow or wide bin width range when presenting graphical presentations of our RD design, the literature recommends conducting both informal and formal tests to find the most appropriate bin width (Jacob et al., 2012). Accordingly, we plot eligibility against the market capitalization of all observations in our sample using different bin widths, ranging from 0.5 to 5. Next, we visually inspect and compare each of the plots to determine whether we observe any significant changes in the relationship between the two variables. These figures are attached in section C2 of the appendix.

The relationship between our outcome and running variable is largely unaffected when we alter the bin width, indicating that all three bin widths pose as equally appropriate options. Based on these results, we deem the informal test to be sufficient in determining the optimal bin width, using a bin width of 2 in the further analyses.

7.4.2 Exogeneity in the cut-point

Assessing the internal validity it also important to determine whether the running variable, or cut-point, could have been manipulated. We have not been able to establish exactly why the cut-point is set at $25 million in market capitalization, but believe that this threshold is set somewhat arbitrarily.

Moreover, we assume that firms are unable to precisely determine their own market capitalization with respect to the threshold set by Robinhood. The rationale is that market capitalization is dependent on the contemporaneous demand among investors in the market, preventing complete firm control over this measure. In accordance with the research of Lee and Lemieux (2010), we therefore assume that stocks just above and below the cut-point are *"as if randomly assigned"*. This randomization implies that the variation in the treatment near the threshold is randomized as though from a randomized experiment, allowing us to interpret the gap at the threshold intersect as the causal effect of fractional share trading on retail investors' preferences.

7.4.3 Sharp versus Fuzzy-regression

Optimally, we would also control for *"no-shows"* or *"crossovers"*. A *"no-show"* in our sample would be any observation which is not traded fractionally despite being eligible, while a *"crossover"* would be an ineligible stock being traded fractionally. We are however unable to precisely observe if such instances occur, as Robinhood has no public information about their rigidness to these rules. Nonetheless, we choose to believe Robinhood's statement, and have as such constructed the RD to be sharp by design.

7.4.4 Comparison of non-outcome variables by propensity score matching

We also conduct an analysis of the relationship between the non-outcome variables and the running variable to ensure that neither of our covariates are affected by the treatment. To determine the comparability of eligible and ineligible stocks near the threshold, we gather data on several firm-specific characteristics and run two sample t-tests for equality in means. Due to the very nature of lottery-type stocks being rare, and also to the FST-service being fairly new, our data sample is restricted with respect to size. With a small sample, group means, and also the estimated p-values of the associated t-tests, are sensitive to extreme observations. To enhance accuracy and mitigate potential bias in the estimation of group means, we therefore use the method of propensity score matching prior to conducting the two-sample t-tests.

In addition to the characteristics previously mentioned, we include four valuation multiples: enterprise value (evm), price-earnings (pe_inc), price-sales (ps) and price-to-book (ptb). We also include dividend payout ratio (drp) and three capital structure ratios: debt over EBITDA ($debt\_ebitda$), cash over debt ($cash\_debt$) and debt over equity (de_ratio). The results are reported in table 7.2 below.

Table 7.2: Comparison of group means in non-outcome variables

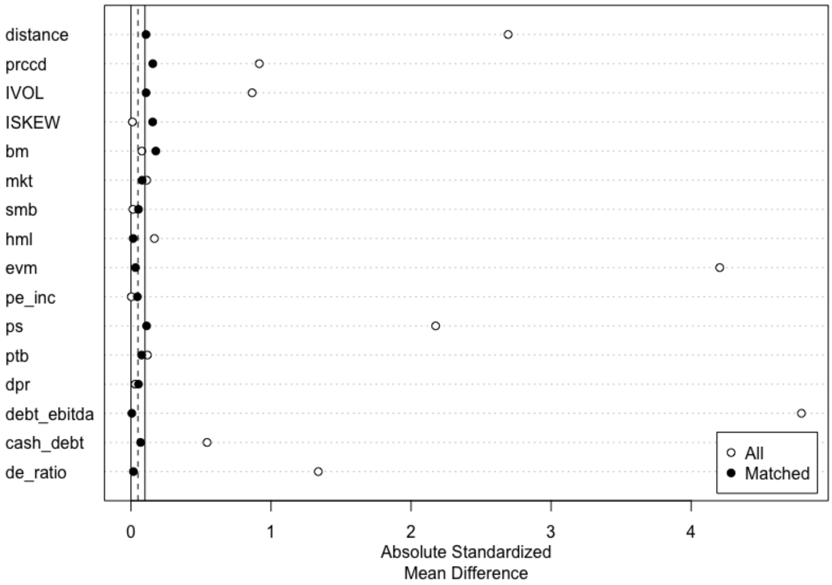
This table reports monthly means of firm characteristics of stocks traded on the Robinhood-platform across our sample period. To compare equality near the threshold, eligible and ineligible stocks are matched based on propensity scores. Columns 0 and 1 refer to the control and the treated group respectively, SMD reports the standard mean differences between the matched samples, while p-value shows the results of the two-sample t-test of equality in means.

Measure (mean(sd))	0	1	p-value	SMD
N	328	328		
Price	8.65 (8.59)	12.62 (8.27)	<0.001	0.471
IVOL	3.04 (1.35)	2.55 (1.43)	<0.001	0.347
ISKEW	1.07 (1.02)	1.14 (1.15)	0.374	0.069
bm	1.65 (1.19)	1.81 (1.62)	0.136	0.117
mkt	0.25 (5.30)	-0.37 (5.54)	0.138	0.116
smb	-2.11 (2.11)	-2.06 (1.94)	0.763	0.024
hml	-0.29 (1.22)	-0.44 (1.53)	0.163	0.109
evm	11.38 (11.26)	11.05 (9.83)	0.693	0.031
pe_inc	17.60 (18.18)	17.56 (18.81)	0.982	0.002
ps	0.86 (0.60)	0.90 (0.75)	0.466	0.057
ptb	0.80 (0.39)	0.91 (0.40)	<0.001	0.279
dpr	0.79 (1.09)	0.70 (1.14)	0.292	0.082
debt_ebitda	3.17 (11.87)	4.08 (9.32)	0.278	0.085
cash_debt	0.30 (0.29)	0.25 (0.34)	0.018	0.185
de_ratio	0.79 (2.75)	1.79 (3.35)	<0.001	0.326

We use strict requirements for matching to ensure firms are matched on equal terms to avoid large distances between observations. The labels of *1* and *0* refers to the treated and control observations, respectively, in which treated observations alludes to the eligible stocks, and vice versa for the observations in the control group. *SMD* represents the standard mean difference after matching, while *p-value* shows the results of the Welch two sample t-tests for equality in group means.

In table 7.2, we observe significant differences in group means between some of our non-outcome variables. However, observing some inequalities in group means across a broad range of firm-characteristics is not surprising, in particular when we consider the small size of our data sample. We also find solace by observing the assessment of standardized mean differences before and after matching in figure 7.2 below. In the following, we therefore assume that the equality of stocks near the threshold is acceptable enough for us to draw inference from our sample. For closer analysis, distribution of propensity scores and histograms for the matched and unmatched samples, we refer to section C3 of the appendix.

Figure 7.2: Standardized mean differences of non-outcome variables in the RD design, before and after propensity score matching



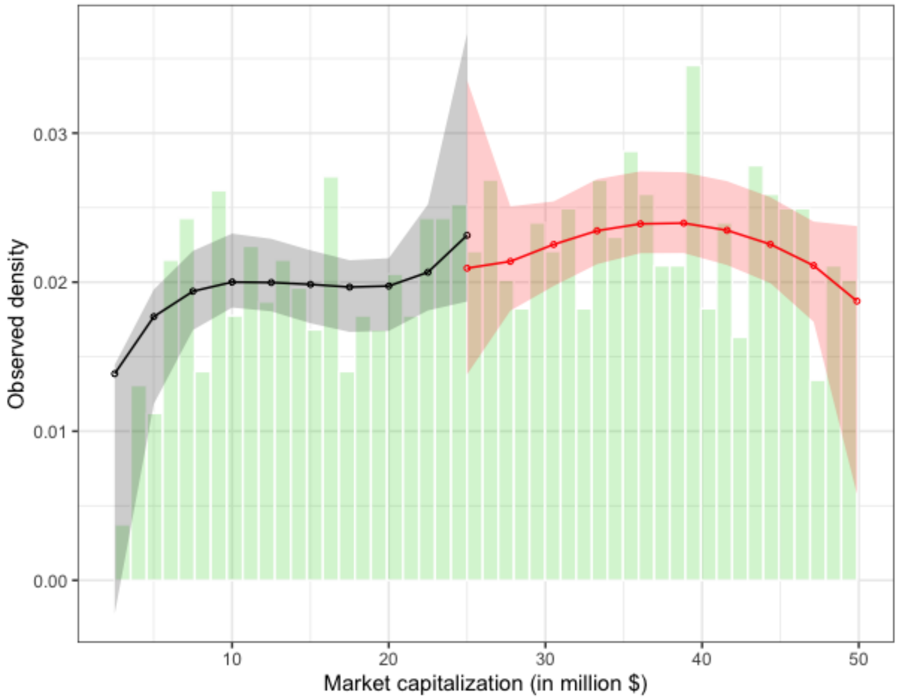
This figure shows the standardized mean differences of all non-outcome variables before and after propensity score matching. Black and white dots refer to the matched and unmatched observations respectively.

7.4.5 Density of the running variable

As eligibility could provide certain benefits, we examine whether there are any significant discontinuities at either side of the market capitalization threshold. If such discontinuities exist, it would indicate that the observed values of market capitalization, or even the placement of the cut-point itself, have been subject to some degree of manipulation. The common approach is to plot the observed frequency of market capitalization at each point along the x-axis, as reported in the figures of section C2 in the appendix. To formally test this however, we also conduct a density test by method of McCrary (2007).

In the McCrary density test, we run two separate local linear regressions at either side of the market capitalization threshold. The regressors are the midpoint values of each running variable bin, while the frequency of each bin is the outcome. Due to our relatively small stock sample, we specify a broad bandwidth of 25. By observing the fitted lines and their associated confidence intervals depicted in figure 7.3 – in which the estimated p-value for the size of the overlap equals 0.78 – we once more reject manipulation of the running variable.

Figure 7.3: Density plot of the RD running variable



This figure shows the density of the running variable, represented by the observed values of market capitalization in our sample. Rather than reporting the frequency of each observed value, market capitalization along the x-axis are grouped into bins, before density is measured for each bin. The bold line represents the t-tests of difference in means at either side of the cut-point of $25M, while the shaded area represents the confidence intervals of the t-tests. The black and red line refers to observations below and above the cut-point, respectively.

In conclusion, after confirming exogeneity in the cut-point, rejecting manipulation in the running variable, examining the non-outcome variables and establishing that the treatment is *"as if randomly assigned"*, we conclude that the internal design is valid.

7.4.6 Functional form

The decision to apply linear regression in our RD design is a result of careful consideration of the relationship between the outcome and running variable, absent of treatment. The conventional approach is to test a variety of functional forms and choosing the specification yielding the best fit (Cook, 2008) (van der Klaauw, 2008). With a limited data set, we choose to follow the recommendations of Jacob et al. (2012), who emphasize that for relatively small data sets, the parametric estimation approach is best suited. In doing so, we are sacrificing potentially unbiased estimates, in exchange for making full use of the limited data we have available.

7.4.7 Model critique

Lastly, we evaluate the precision of the estimates obtained from our RD design. At most, our sample consists of only 552 stock observations, posing a threat to the validity of the conclusions we can draw from the results. Since the power to detect causal effects of an RD design is weaker than that of a comparable randomized experiment, one generally demands a larger sample size. Optimally, we should thus have more than twice the number of observations than that in a randomized experiment to achieve the same precision (Jacob et al., 2012). As the size of the sample is imperative in achieving precise estimates, we choose to lay most of our trust in model (1) from table 7.1.

Another concern for our RD design is the generalizability of our estimates. By design, our model only measures the local effects of tradability on stock holdings around the cut-point of the running variable. If strictly interpreted, our findings does not provide any information about the size of this effect for stocks with a market capitalization further away from the $25 million threshold. In contrast, Lee and Lemieux (2010) offer a more expansive interpretation, highlighting that (1) control over the running variable in RD designs are typically imprecise and (2) the outcome variable usually contains considerable occurrences of random errors. As both these incidents can be argued to occur in our RD design, they induce heterogeneity to the cut-point, and increase the generalizability of our findings. As previously argued, stocks in our RD design are subject to an *"as if random"* assignment above and below the threshold. Following the insights offered by Lee and Lemieux (2010), we thus argue that our findings could also apply to a broader population of stocks.

7.5 Difference-in-Difference

To further test the effects of fractional share trading on retail investors' behaviour, we use a Difference-in-Difference approach. This design allows us to analyse the difference in the average number of Robinhood users holding low- and high priced stocks before and after the introduction of fractional share trading. We also conduct a subgroup analysis, in which we follow the procedure of (Brookes et al., 2004) and interact our subgroup of highly return skewed stocks with our treatment group, in order to estimate the difference between the treatment effect on high- and low return skewed stocks.

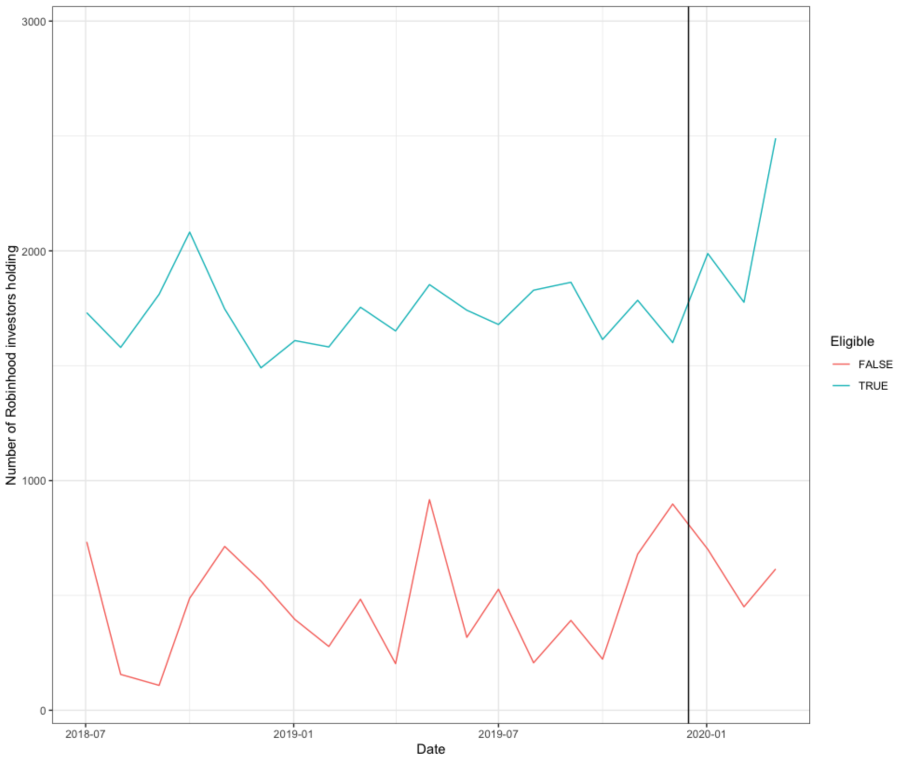
To examine this effect, we exploit the $1 cut-point to define treatment. We also use this threshold to categorize stocks as either low- or high-priced. However, eligibility for FST is also conditional on a market capitalization rule. As we want to isolate the effect of eligibility with respect to price, we only include stocks with a reported market capitalization equal to or above the $25 million eligibility threshold. Our treatment group is the remaining stocks with a price above $1 (*high-priced and eligible)* and our control group is the remaining stocks with price below $1 (*low-priced and ineligible).*

As we are unable to observe the pace at which Robinhood actually implements fractional share trading for stocks crossing the $1 threshold, we limit our DiD analysis to a two-period model to reduce the risk of having non-compliance in our data set. This decision is rooted in the observation that stocks routinely move back and forth between the $1 threshold across the full sample period. If not accounted for, we run the risk of having noncompliance in our data set, which will reduce the credibility of the estimated treatment effect in the regression. Through our two-period DiD we also ensure that we observe the same treatment and control groups at common dates both before and after treatment. All observations prior to treatment are observed at the 1st of December 2019, while we observe all stocks after treatment at February 1st, 2020. The choice of February 1st is made to allow for retail investors to adapt to the new service, and to account for potential errors preventing stocks to be traded fractionally in the early phase of the introduction. Moreover, as we are only interested in the aggregate preferences of retail investors and use a two-period model, we choose to refrain from estimating the regression with fixed effects.

To dentify appropriate counterfactuals, we propensity match the treated stocks with the non-treated stocks, based on book-to-market ratio, Standard Industry Classification Code (SICC) and market capitalization. The propensity score is calculated using a logit model to predict the probability of being traded fractionally – yielding an estimated counterfactual stock for each eligible stock. We expect these two groups to have a similar trend in the number of users owning these stocks, absent of treatment. This crucial assumption allows us to draw a conclusion on whether there is an effect of being traded fractionally – as any further difference between the retail investor stock holdings of the two groups after FST must then be caused by the treatment. When we use propensity matching to identify counterfactuals, we are left with a control group consisting of only 62 observations. This observation is however not too surprising, as there are only a reported 226 stocks with a reported price less than $1 on the NYSE, Nasdaq and AMEX per December 2020, according to MarketWatch's stock screener (MarketWatch, 2020). The distribution plots and histograms of propensity score mathcing is enclosed in section D1 of the appendix.

As the parallel trend assumption is the most crucial assumption for our DiD-model to be valid, we plot the observed aggregate number of Robinhood investors' holding both treated and non-treated stocks, before and after the introduction of FST. In an attempt to capture the trend prior to treatment, we choose to include stock observations from 2018. Even though there are some statistical tests for the parallel trend assumption, these are critiqued as they are only able to measure this trend in the pre-treatment period, and there is no commonly accepted procedure to be sure whether the common trend still holds after the treatment period (Khan-Lang and Lang, 2018). One reason why a parallel trend might not hold in our data is if the proportion of price-sensitive investors at Robinhood is changed throughout the period. Then the low-priced stocks might exhibit a trend caused by inflow of disproportionately price-sensitive investors, rather than an increased preference for low-priced stocks. To our knowledge, however, Robinhood's target customer group did not change throughout this period, which leads us to believe that the parallel trend assumption holds. We plot the trends to visually assess of the common trend assumption in figure 7.4.

Figure 7.4: Parallel trend assumption



This figure shows the observed time-series of the average number of Robinhood investors holding both treated and non-treated stocks, before and after the introduction of FST. The vertical line marks the date of the introduction.

Inspecting figure 7.4, the treatment and control group seem to exhibit similar behaviour on average, suggesting that both groups likely had parallel trends prior to the treatment. Moreover, it seems to be a considerable divergence in Robinhood investors' holdings immediately after the introduction of fractional share trading on December 16th, 2019. We also note that the average number of retail investors in eligible stocks, i.e. stocks with price above $1, is noticeably higher. We interpret this as a sign that the average retail investor price threshold is above $1.

In equation 7.1 below, we report our Difference-in-Difference model, in which both *Eligible* and *After* are binary variables. *Eligible* is equal to 1 for stocks eligible for FST, defined by stock price equal to or above $1. *After* takes the value of 1 if the date is after December 16th, 2019, and zero otherwise. *log(Mktcap)*, *log(Price)* and *log(MAX)* are included as control variables, and are log-transformed to mitigate the effect of potential outliers in their positively skewed distributions. In the same spirit, we also log-transform our dependent variable, *Users holding*. With a partial log-log regression model, the interpretation of the estimated coefficients of the log-transformed variables changes, in which a 1% increase in our explanatory variables translates to the approximate percentage change in *Users holding*. For the coefficients of the non-transformed variables however, a 1 unit increase yields the approximate percentage change in *Users holding*.

$$log(Usersholding)_{i,t} = \alpha_{i,t} + \beta_1 Eligible * After_i + \beta_2 Eligible_{i,t} + \beta_3 After_i + \tag{7.1}$$

$$\beta_4 log(Mktcap)_{i,t} + \beta_5 log(Price)_{i,t} + \beta_6 log(MAX)_{i,t} + \epsilon_{i,t}$$

The DiD estimate, given by the coefficient of the regular interaction term *Eligible*After*, can be interpreted as the causal effect on stock ownership when stocks can be traded fractionally. The reference category for this regression is thus the number of users holding stocks with price less than $1 before fractional share trading was implemented.

In order to investigate our *Wealth Constraint Hypothesis*, we also run a subgroup analysis of the stocks with return skewness above median by interacting this group with the DiD estimator from equation 7.1. Thus, for our hypothesis to be correct, the causal effect on the number of Robinhood holding stocks that are being traded fractionally should be less than for stocks with low return skewness. The final model is therefore extended to:

$$log(Usersholding)_{i,t} = \alpha_{i,t} + \beta_1 Eligible * After * HighSkew_i + \beta_2 Eligible * After_i + \tag{7.2}$$

$$\beta_3 Eligible_{i,t} + \beta_4 After_i + \beta_5 log(Mktcap)_{i,t} + \beta_6 log(Price)_{i,t} + \beta_7 log(MAX)_{i,t} + \epsilon_{i,t}$$

Table 7.3: Difference-in-Difference regressions on retail investor demand after the introduction of Fractional Share Trading

This table reports the DiD panel regressions. The number of Robinhood users holding a stock is log-transformed and included as the dependent variable. Eligible indicates whether a stock is eligible for FST, After takes the value of 1 if the stock is observed after the introduction of FST, while HighSkew indicates the stocks with above median return skewness. Model (1) captures the causal effect of fractional share trading on retail demand for high-priced, eligible stocks, while model (2) in addition controls for high return skewness. The log of market cap, price and past maximum return are included as controls. We use clustered standard errors and t-statistics are reported in parentheses.

	Dependent variable:	
	log(Users holding)	
	(1)	(2)
Eligible*After*HighSkew		-0.702
		(-0.909)
Eligible*After	0.037	0.503
	(0.135)	(0.773)
Eligible*HighSkew		0.551
		(0.752)
After*HighSkew		0.238
		(0.401)
Eligible	0.469	0.099
	(1.164)	(0.120)
After	0.361	0.163
	(1.523)	(0.310)
HighSkew		-0.312
		(-0.627)
log(Mktcap)	0.753^{***}	0.754^{***}
	(8.146)	(8.074)
log(Price)	-1.129^{***}	-1.127^{***}
	(-7.355)	(-7.394)
log(MAX)	0.017	0.027
	(0.140)	(0.172)
Constant	3.314^{***}	3.522^{***}
	(4.944)	(3.475)
Fixed Effects	No	No
Observations	124	124
R^2	0.641	0.646
Adjusted R^2	0.623	0.615
F Statistic	34.848^{***} (df = 6; 117)	20.647^{***} (df = 10; 113)
Note:		*p<0.1; **p<0.05; ***p<0.01

If we are to believe the DiD-estimator given in model (1), expressed by the term *Eligible*After*, the introduction of fractional share trading will causally increase the number of users holding high-priced stocks by 3.7 percent on average. This result is not close to significant, so this result should not be interpreted as anything else than a trend in the data. When we extend our model to also include a subgroup analysis of highly return skewed stocks in model (2), this coefficient changes significantly. We interpret this as a sign of the inherent uncertainty of the coefficient of model (1). Consequently, reporting an exact magnitude of the treatment effect is speculative, and we therefore limit ourselves to noting that the trend is positive in both models.

The main coefficient of interest in model (2) is however the double interaction term, *Eligible*After*HighSkew*. The coefficient is negative and insignificant, and expresses the difference in the magnitude of the effect of fractional share trading between stocks with highly skewed returns and stocks with low return skewness. The coefficient shows that there is about a 50% difference between these two stock groups, in which the stocks with above median return skewness are at the negative side of the difference. Nevertheless, these coefficients are insignificant, and we cannot draw any conclusion based on this analysis. These results are highly significant. The trend is however in favour of our *Wealth Constraint Hypothesis* - retail investors shy away from stocks with highly skewed returns when their wealth constraints are reduced.

Finally, we note that the coefficients of *log(Mktcap)*, *log(Price)* and have the expected signs. Thus, a 1% increase in market capitalization is associated with about a 0.75% increase in the number of users holding a stock. Also, a 1% increase in the number of users holding is associated with a decrease of about 1.12% in the number of users holding a stock on average. We also note that the estimated coefficient of *log(MAX)* is positive but insignificant. This result is however not surprising, as the model accounts for the time-dimension through the *After* dummy.

7.5.1 Model critique

There are several weaknesses in the estimates obtained in the previous section. First, the is little economic significance of a threshold set at $1 for FST eligibility. Whereas we do believe retail investors are sensitive to the nominal stock price, we also believe that the actual significance on retail investor holdings is more profound when prices are further away from $1. As the descriptive statistics in table 7.4 shows, the median price of the treated stocks is $4.23, which is arguably too small to make any economic impact on retail investor behavior. This view is further supported by the parallel trend plot in figure 7.4, in which stocks with price above $1 also have considerably more users holding. The table also shows that the distribution of stock prices is skewed towards relatively low prices. If the threshold was rather set at a price with more economic significance, we expect the observed effects to be more pronounced. A solution to this problem might be to remove all stocks above a price threshold (say $10) and run the regression again. However, this modification would likely weaken the matching of the groups, as one can argue that there are fundamental differences between low- and high-priced stocks.

Table 7.4: Descriptive price statistics of treated and control stocks

This table reports a selection of price characteristics for treated and control stocks.

Price	Treatment	Control
Mean	10.59	0.73
5th-quantile	1.27	0.42
25th-quantile	2.97	0.57
50th-quantile	4.23	0.77
75th-quantile	7.44	0.89
95th-quantile	30.53	0.97
Min	1	0.25
Max	180.79	0.98
Number of observations	62	62

The number of observations is also a cause of concern for both models, but perhaps more so when conducting our subgroup analysis. The subgroup analysis requires the data to be subdivided into even smaller groups (*low* versus *high*), which reduces the ability of the regression to detect a treatment effect (Brookes et al., 2004). Thus, with a small data set, our subgroup analysis is more likely to report false negatives, and is therefore less reliable than the analysis in model (1).

In sum, we consider this model too unreliable to conclude on any findings, due to a small data set, and to the fact that the economic significance of the $1 threshold is arguably too small to detect any treatment effect. Nevertheless, we note that the trend is in favour of our *Wealth Constraint Hypothesis*.

7.6 Correlational analysis of idiosyncratic skewness and retail investment

Finally, we run multiple linear regression to further test the relationship between idiosyncratic skewness and retail investment. In line with the *Wealth Constraint Hypothesis*, and our findings in the analyses above, this relationship should be insignificant when controlling for stock tradability. Hence, given the opportunity to trade a wider spectrum of stocks, retail investors should shy away from stocks with highly skewed returns.

To test this prediction, we redefine our model specification, including idiosyncratic skewness as our dependent variable. The time span covered in this model is the period after fractional share trading was implemented. We use *mktcap_ centered*, *EligibleDummy*, *Users holding* and an interaction term of the *EligibleDummy* and *Users holding* are included as regressors. EligibleDummy is equal to 1 for all stocks that are traded fractionally, and zero otherwise.

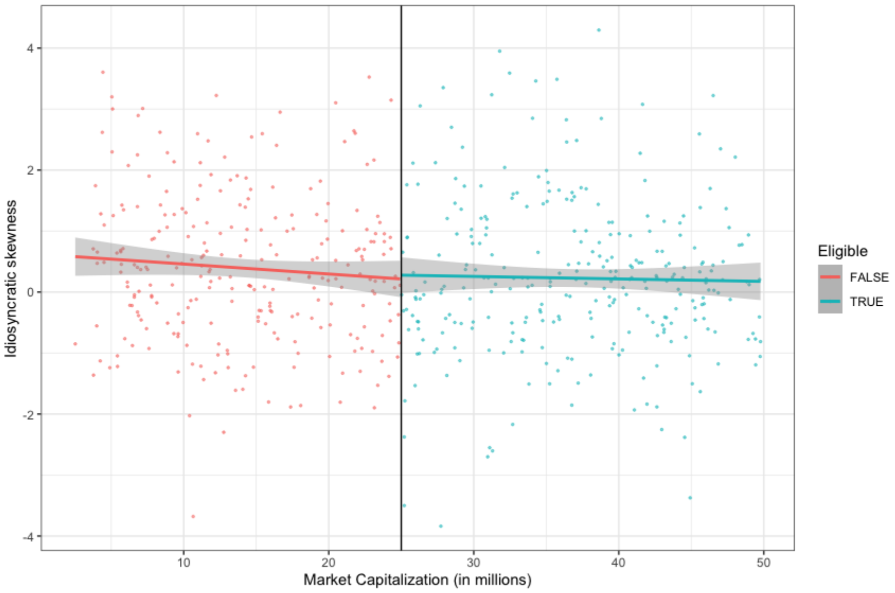
Table 7.5: OLS on idiosyncratic skewness and stock tradability

This table reports the OLS regression estimates for idiosyncratic skewness, proxying as a lottery characteristic, on market capitalization (the running variable), number of Robinhood investors holding, an eligibility dummy indicating stock tradability of fractional share trading, and an interaction term between Robinhood investors holding and the eligibility dummy. Market capitalization is centred to indicate the distance of each observation from the cut-point. T-statistics are reported in parentheses.

	Dependent variable:
	Idiosyncratic skewness
EligibleDummy	0.171
	(0.772)
Users holding	0.0002***
	(2.829)
mktcap_centered	−0.011
	(−1.424)
EligibleDummy*Users holding	−0.0001
	(−0.940)
Constant	0.113
	(0.871)
Model	OLS
Observations	552
Adjusted R^2	0.020
Note:	*p<0.1; **p<0.05; ***p<0.01

As we see in table 7.5, *Users holding* is positive and significant, and we can interpret the estimated coefficient of 0.0002 as the average increase in idiosyncratic skewness of an ineligible stock's return that is associated with one additional investor in that stock. The relationship between *Users holding* and *Idiosyncratic skewness* disappears when we control for tradability through the interaction term *EligibleDummy* Users holding.* This finding strengthens our argument that, when controlling for tradability, retail investors do not actively pursue stocks with skewed returns. The trend found in the RDD and DiD analyses is also more convincing when seen in relation to this finding.

Figure 7.5: Relationship between idiosyncratic skewness and eligibility



This figure shows the observed idiosyncratic skewness of stock returns above and below the eligibility threshold of $25M, using a bandwidth of 25. The vertical line marks the cut-point.

In order to verify our choice of functional form, we have also plotted the observations of idiosyncratic skewness against the running variable. Similarly to the number of Robinhood investors holding plotted in 7.1, the linear regression appears to be the best fit as the observed idiosyncratic skewness are homogenously distributed across the plot.

7.7 Generalizability of findings

A considerable cause of concern for the generalizability of our results is the representativeness of Robinhood investors for retail investors as a whole. First, Robinhood's investors have been described in the media as being young millennials, as their average user per 2020 is 31 years old (Rooney, 2020). Additionally, 78% of Robinhood's users were under the age of 35 in 2017, which is a critical drawback of our analysis (Harris, 2017). Moreover, Robinhood investors are being described as more risk-seeking, as the Robinhood app is a gamified experience - for example displaying confetti when users purchase stocks (Ingram, 2019). Thus, many believe that gamblers turn to Robinhood as a complement to sports betting or the like (Maggiulli, 2020). These issues pose a threat to the generalizability of Robinhood investors' preference for return skewness, which is a key feature of our lottery-like stocks. Additionally, young investors with low portfolio value are more likely to invest through mobile apps (Lin and Mottola, 2019), which makes this group more likely to exhibit capital constraints when investing in the stock market. As a conclusion, we note that the Robinhood investors' preferences for lottery-like stocks might be more pronounced than average. However, we would not go as far as detaching this group from other retail investment brokers as a whole and argue that the mechanisms our analyses show are likely to persist to some extent if one were to expand the sample to other retail brokers as well.

7.8 Implications of findings

From our analysis, we find some evidence, although not fully conclusive, suggesting that retail investors are overly concentrated in lottery-type stocks due to wealth constraints. Expanding on our findings, we argue that fractional share trading, through providing retail investors access to a broader investment universe, reduces friction and increases efficiency in the capital markets. In particular, we would expect the pronounced overpricing in lottery-type stocks to gradually disappear, as retail investors are able to shift towards stocks with higher prices and lower return skewness. Accordingly, we expect the existence of negative lottery premiums to be reduced if fractional share trading is offered to a wide spectrum of investors. As retail investors, and Robinhood investors especially, account for a small percentage of the total market, we are however uncertain of the effect on efficiency for the market as a whole.

Removal of the price barrier will also reduce firms' incentives to perform active stock price management. Stock price management, for example through stock splits, usually appear as bad value propositions, incurring costs while providing no added fundamental value. As previously mentioned however, proponents of the *Marketability Hypothesis* argue that firms should actively keep nominal stock prices low, as part of a marketability scheme to attract certain investors. While such a hypothesis could hold, the introduction of fractional share trading reduces managers' incentives for such activities. Thus, firms have less incentives to target investors through low nominal prices if fractional share trading is implemented on a wide basis.

We would however like to emphasize that there are considerable uncertainty associated with our findings, both due to the limited sample size, but also because the FST-service is relatively young. As a consequence, the estimates and results obtained from both our RD and DiD designs must be interpreted with caution. Thus, the effect of fractional share trading on retail investor demand for low- and high-priced stocks should be retested when more data is available. Future research should also examine whether the observed effects are consistent across different brokerage platforms.

8 Conclusion

We find evidence, although not conclusive, that retail investors on aggregate show a preference against lottery-type stocks once we control for stock tradability through the Fractional Share Trading service. This trend is documented using the methodologies of Regression Discontinuity Design, Difference-in-Difference and standard OLS. Our findings give support to the proposed *Wealth Constraint Hypothesis* - in which we expect retail investors to drive the negative lottery premiums in the stock market, as their wealth constrained investment universe is inherently more lottery-like. Although the results from both experimental designs are subject to considerable uncertainty, we observe a common trend in all three models: *retail investors shift towards stocks with less skewed returns once their capital limitations are reduced.*

References

Baker, H. K. and Gallagher, P. L. (1980). Management's view of stock splits. *Financial Management*, pages 73–77.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *National Bureau of Economic Research.*

Barberis, N. and Huang, M. (2007). Stocks as lotteries: The implications of probability weighting for security prices. *National Bureau of Economic Research.*

Benartzi, S., Michaely, R., Weld, W. C., and Thaler, R. H. (2009). The nominal share price puzzle. *Journal of Economic Perspectives 23 (2)*, pages 121–142.

Birru, J. and Wang, B. (2016). Nominal price illusion. *Journal of Financial Economics.*

Brookes, S., Egger, M., Mulheran, P., Peters, T., Smith, G., and Whitely, E. (2004). Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of Clinical Epidemiology 57(3)*, pages 229–236.

Chen, M., Chen, S., Yen, M., and Shen, Y. (2008). Lottery premium in the taiwan stock market. *Asia Pacific Management Review 13(2)*, pages 545–556.

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression discontinuity design in psychology, statistics and economics. *Journal of Econometrics 142(2)*, pages 636–654.

Dahr, R., Goetzmann, W. N., Shepherd, S., and Zhu, N. (2004). The impact of clientele changes: Evidence from stock splits. Technical report.

Downs, T. W. and Wen, Q. (2001). Is there a lottery premium in the stock market? *The Journal of Portfolio Management 28(1)*, pages 112–119.

Dyl, E. A. and Elliott, W. B. (2006). The share price puzzle. *The Journal of Business*, pages 2045–2066.

Easley, D., O'Hara, M., and Saar, G. (2001). How stock splits affect trading: A microstructure approach. *Journal of Financial and Quantitative Analysis 36*, pages 25–51.

Fama, E. F. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*, pages 3–56.

Fama, E. F. and MacBeth, J. D. (1973). Risk and return: Some empirical tests. *Journal of Political Economy 81*, pages 606–636.

Green, T. C. and Hwang, B. (2009). Price-based return comovement. *Journal of Financial Economics 93 (1)*, pages 37–50.

Han, B. and Kumar, A. (2008). Retail clienteles and the idiosyncratic volatility puzzle. *McCombs Research Paper Series.*

Harris, A. (2017). How brokerage app robinhood got millenials to love the market. https://www.fastcompany.com/40437888/how-brokerage-app-robinhood-got-millennials-to-love-the-market.

Harvey, C. R. and Siddique, A. (2000). Conditional skewness in asset pricing tests. *Journal of Finance 55*, pages 1263–1295.

Ibragimov, R. and Muller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28(4)*, pages 453–468.

Ince, O. S. and Porter, R. B. (2006). Individual equity return data from thomson datastream: Handle with care! *Journal of Financial Research 29*, pages 463–479.

Ingram, D. (2019). Designed to distract: Stock app robinhood nudges users to take risks. https://www.nbcnews.com/tech/tech-news/confetti-push-notifications-stock-app-robinhood-nudges-investors-toward-risk-n1053071.

Jacob, R., Zhu, P., Somers, M. A., and Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC*.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*.

Khan-Lang, A. and Lang, K. (2018). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business Economics Statistics*.

Kumar, A. (2009). Who gambles in the stock market? *The Journal of Finance*.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity design in economics. *Journal of Economic Literature 48*, pages 281–355.

Lin, J. and Mottola, G. (2019). Investors in the united states: A report of the national financial capability study. https://www.usfinancialcapability.org/downloads/NFCS_2018_Inv_Survey_Full_Report.pdf.

Maggiulli, N. (2020). No, robinhood traders aren't affecting the stock market. https://marker.medium.com/no-robinhood-traders-arent-affecting-the-stock-market-8758af5ad319.

MarketWatch (2020). https://www.marketwatch.com/tools/stockresearch/screener/.

McCrary, J. (2007). Manipulation of the running variable in the regression discontinuity design: A density test. *National Bureau of Economic Research 334*.

Norland, E. (2020). Why large caps have done better than small caps. https://www.cmegroup.com/education/featured-reports/why-large-cap-stocks-have-outperformed-small-caps.html.

Robinhood (2020). https://robinhood.com/us/en/support/articles/fractional-shares/.

Robintrack (2020). https://robintrack.net/.

Rooney, K. (2020). Fintech app robinhood is driving a retail trading renaissance during the stock market's wild ride. https://www.cnbc.com/2020/06/17/robinhood-drives-retail-trading-renaissance-during-markets-wild-ride.html.

Schultz, P. (2000). Stock splits, tick size, and sponsorship. *Journal of Finance 55*, pages 429–450.

Switzer, L. N. (2010). The behaviour of small cap vs large cap stocks in recessions and recoveries: Empirical evidence for the united states and canada. *North American Journal of Economics and Finance.*

U.S. Treasury (2020). https://home.treasury.gov/.

van der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour 22(2)*, pages 219–245.

Vokatá, P. (2012). Gambling in stock markets: Empirical evidence from europe. *Faculty of Social Sciences, Institute of Economic Studies. Charles University in Prague.*

Wharton Research Data Services (2020). https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/compustat/.

Appendix 1: Data

A1 Data screening, cleaning and matching

For each stock we obtain gvkey (firm identifier), unadjusted daily prices (mean absolute value of the bid-ask spread), a daily adjustment factor, a total daily return factor, exchange- and industry codes and number of shares outstanding from the Compustat database (Wharton Research Data Services, 2020). All stock prices are denominated in dollars. Further, we also retrieve balance sheet data, including shareholder equity, total equity, total assets, total liabilities, tax investment credits, deferred taxes and preferred stock redemption-, liquidity- and par value. We also download daily popularity data on the number of Robinhood investors holding U.S. stocks from Robintrack (2020).

Daily returns are computed by applying the Compustat convention of dividing unadjusted closing prices with a daily adjustment factor, before multiplying with a total daily return factor. Inspired by Ince and Porter (2006), we omit any daily returns above 400% and below -85% to adjust for outliers and errors in the data set. We also set both *Rt* and *Rt-1* to NA if *Rt* or *Rt-1* is greater than 300% and if *(1+Rt)-(1+Rt-1)-1* is less than 50% (indicating extreme reversal). Note that subscript t in this case refers to monthly returns.

Excess returns, both daily (d) and monthly (t), are calculated as returns on the market minus the appropriate risk-free rates. As a proxy for the risk-free rates, we use the time series of 1-month T-bill rates from the U.S. Department of the Treasury (U.S. Treasury, 2020). Since the risk-free rates are reported on a monthly basis, we convert the rates from monthly to daily by the common convention:

$$rf_d = (1 + rf_t)^{\frac{1}{d}} - 1$$

where d refers to number of days in month t

Market capitalization is computed as the stock price times the number of shares outstanding. In order to compute the book-to-market ratio, we derive the book equity value from the financial data obtained from Compustat. We apply the following formula to back out the book equity value of each stock:

Book Equity=(total equity + total assets + tax credits + deferred taxes) – (total liabilities + preferred stock redemption value - preferred stock liquidity - preferred stock par value)

A2 Variable explanation

After downloading, screening, cleaning and matching the necessary data, we are left with a data set on U.S. common stocks across the time period of 2000 to 2020. The data set is balanced, and comprises the following variables:

gvkey: firm identifier

date: year-month-day

reference_ date: date for Fama-French portfolio sorting

exchange: exchange code

ret: return on market

rf: proxy for risk-free rate

R: excess return

prccd: unadjusted stock price

mktcap: market capitalization, Fama-French size proxy

bm: book-to-market ratio, Fama-French value proxy

mktcap_ weight: market capitalization on June each year

Appendix 2: Replication of lottery stock premiums

This part provides additional analyses related to the replication efforts described in section 6. As we question the role of idiosyncratic volatility as a lottery characteristic, we first run cross-sectional Fama and MacBeth (1973) regressions on two separate lottery portfolios. The first portfolio is sorted using the traditional lottery definition of Kumar (2009), while the second portfolio is sorted only on stock price and idiosyncratic skewness. In doing so, we are able to investigate the existence of lottery stock premiums under different lottery stock definitions. We also present an overview of the correlations between the estimated lottery characteristics. Moreover, by taking basis in the lottery portfolio sorted under the alternative lottery stock definition, we investigate portfolio performance and time series effects of the lottery stock premium across our sample period. Lastly, we conduct a robustness analysis to ensure the validity of our results.

B1 Results with traditional and alternative lottery stock definition

In table B1.1, we present the results of the cross-sectional Fama-MacBeth regressions on the two lottery portfolios. The first portfolio is sorted on the three traditional lottery characteristics proposed by Kumar (2009), while the second portfolio constitutes stocks with price below median and idiosyncratic skewness above median. If idiosyncratic volatility is merely a consequence of retail activity, and not a lottery feature, we would expect the lottery premiums to be equally pronounced when we regress both portfolios.

Note that model (2) in table B1.1 is the same as in table 6.1. In this part however, we have altered the variable name to *AlternativeLotteryDummy*. This is done in order to readily compare the premiums from section 6 with the estimated premiums from a lottery portfolio sorted on the three traditional lottery characteristics, represented by *TraditionalLotteryDummy*. The output in table B1.1 report the existence of significantly negative lottery premiums when we regress on both the traditional and the alternative lottery portfolio. The estimated lottery premiums are -0.785 and -0.501, respectively, both significant at the 1% level. In annual terms, this adds to -9.42% and -6,01%.

Table B1.1: Lottery premiums with traditional and alternative lottery stock definition

This table reports the Fama and MacBeth (1973) monthly cross-sectional regression estimates for the value-weighted excess return on both the traditional and alternative lottery-type portfolio. Model (1) is regressed on a portfolio sorted using the traditional lottery-stock definition of Kumar (2009), while model (2) is regressed on the alternative lottery portfolio – sorted on below median price and above median idiosyncratic skewness. Contemporaneous factor betas for the Fama-French risk factors (mkt, smb, hml), size (log of market value), book-to-market and lagged 6-month return are included as controls. T-statistics are reported in parentheses.

	Dependent variable:	
	Value-weighted return	Value-weighted return
	(1)	(2)
TraditionalLotteryDummy	-0.785^{***} (-6.665)	
AlternativeLotteryDummy		-0.501^{***} (-5.386)
mkt	0.936^{***} (5.033)	0.937^{***} (5.008)
smb	-0.284 (-0.805)	-0.309 (-0.874)
hml	0.966^{***} (7.546)	0.977^{***} (7.584)
Log_mktcap	2.103^{***} (16.409)	2.133^{***} (16.333)
bm	0.882^{***} (11.893)	0.883^{***} (11.852)
LaggedR	-0.011^{**} (-2.574)	-0.011^{**} (-2.500)
Constant	13.422^{***} (6.319)	13.465^{***} (6.207)
Model	Fama-MacBeth	Fama-MacBeth
Observations	671,039	671,039
R^2	0.230	0.230
Note:		*p<0.1; **p<0.05; ***p<0.01

In conclusion, we document the existence of negative lottery premiums during the last two decades in the American stock market using two different lottery stock definitions. This finding is also consistent with the previous research on lottery premiums, documented by Kumar (2009). Moreover, and perhaps even more interesting, we find these premiums to be somewhat equally pronounced, and significant, when we construct a lottery portfolio using our alternative lottery definition. Despite only being a correlational analysis, this yields enhanced support to the claim that idiosyncratic volatility is not a lottery feature of attraction, but rather a consequence of retail investor activity. Based on these findings, we therefore take basis in our alternative lottery stock definition in the following replication efforts.

B2 Correlation of lottery characteristics

In spite of our main analysis only making use of idiosyncratic skewness to identify lottery-type stocks, we however compute *IVOL*, using the three-factor model of Fama and French (1993), and *MAX*, using the method of Bali et al. (2011) for our robustness analysis. In table B2.1 we report the correlations between the three lottery characteristics.

Table B2.1: Correlation matrix of lottery characteristics

This figure reports the correlations between the three measured lottery characteristics – idiosyncratic volatility, idiosyncratic skewness and past maximum return observed over the previous month.

	IVOL	ISKEW	MAX
IVOL	1		
ISKEW	0.146	1	
MAX	0.892	0.389	1

We observe a considerable co-movement between idiosyncratic volatility and past *MAX*-returns, with a correlation of just below 90%. This correlation suggests that stocks exerting high idiosyncratic volatility also tend to yield more extreme positive returns. Moreover, we observe a positive correlation of 38.9% between *ISKEW* and *MAX*, which is consistent with the claim that retail investors could use past extreme returns as a proxy for identifying skewness (Kumar, 2009) (Bali et al., 2011). Lastly, the estimated correlation between *IVOL* and *ISKEW* is 14.6%, indicating a weak, but positive relationship between the two measures. This observed lack of co-movement is however consistent with our predictions, as we clearly distinguish the two measures in our analysis – i.e. we view idiosyncratic skewness as a lottery characteristic, while we deem idiosyncratic volatility to be a consequence of retail investor activity.

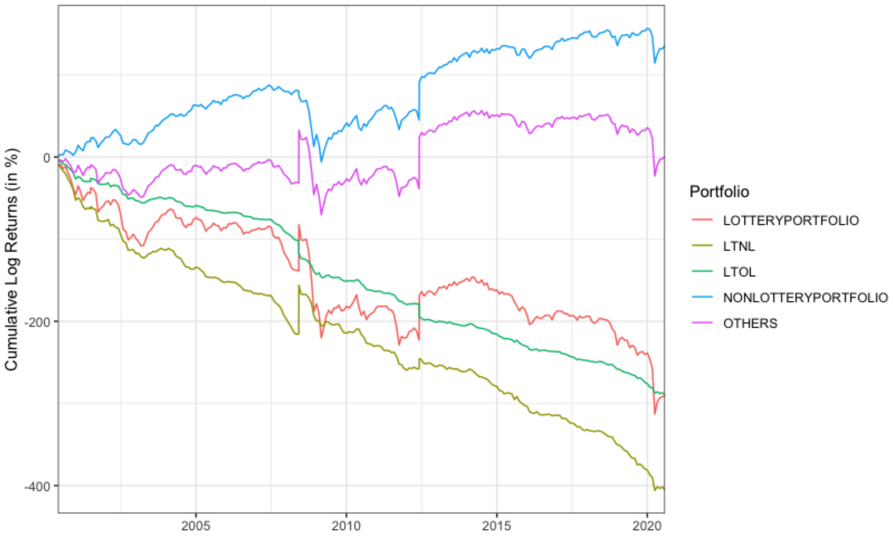
B3 Portfolio performance

To evaluate the performance of the lottery portfolio, we apply portfolio thresholds in similar fashion to the method proposed by Kumar (2009), and form our lottery-type portfolio (*Lottery-Type*) of stocks with below median price *(L)* and above median return skewness *(H)*. Stocks with above median price *(H)* and below median return skewness *(L)* make up the non-lottery-type portfolio (*Non-Lottery-Type*). The remaining stocks are grouped in a residual portfolio labelled *Other*. The descriptive statistics of a selection of firm characteristics across the three portfolios is reported in Table 6.2 and commented on in section 6.3.

In addition to the three portfolios sorted on lottery stock characteristics, we construct two new portfolios in order to measure performance in greater detail. *LT-NL* represent a zero-net-investment portfolio that is long lottery-type stocks and short non-lottery-type stocks. Hence, *LT-NL* is computed by subtracting the monthly value-weighted returns of the *Lottery-Type* portfolio with ones from the *Non-Lottery-Type* portfolio. *LT-OL* is constructed in similar fashion, except we subtract the stocks from the *Other* portfolio.

In Figure B3.1, we compare portfolio performance by plotting the cumulative value-weighted log-returns of the five portfolios across the sample period. We observe a divergence between the lottery- and non-lottery portfolios, in which lottery-type stocks underperform considerably. This difference is captured and portrayed in the *LT-NL*-portfolio, yielding negative cumulative return of roughly 400% in 20 years. Controlling for the validity, we observe dramatic drops in value-weighted cumulative returns across the three sorted portfolios (*Lottery-Type*, *Non-Lottery-Type* and *Other*) in early 2000, 2008 and in 2020 – all consistent with periods of financial crises and distress.

Figure B3.1: Portfolio performance from 2000 to 2020

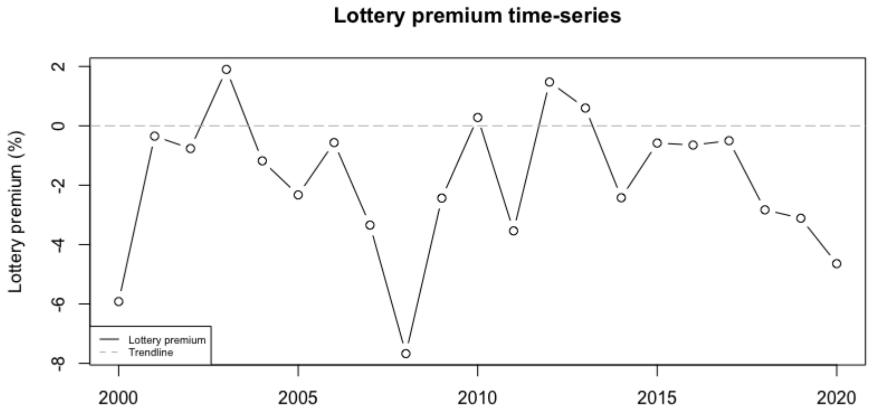


This figure shows portfolio performances across the sample period of the five constructed portfolios described in section B3, measured by cumulative value-weighted log-returns.

B4 Time series effects

In addition to investigating the lottery premiums on an aggregate firm-level, we also observe the development of such premiums in the American stock market across our sample period. By conducting yearly regressions on the alternatively sorted lottery portfolio, we are able to compute the lottery premium time series, as reported in figure B4.1. Plotted against the horizontal line set equal to zero, the lottery stock premium – though exhibiting episodic and spontaneous behaviour – is predominantly negative with a mean of -1.8% Moreover, we observe that the lottery stock premiums are most pronounced in down-markets - i.e. during the financial crises in 2000, 2008 and most recently 2020 - consistent with the findings of Chen et al. (2008) and Vokatá (2012).

Figure B4.1: Time series of lottery stock premiums



This figure shows the estimated annual premiums on the lottery-type portfolio across the sample period of 2000 to 2020. The dashed trend line is set equal to zero for means of comparison, while the bold line and dots represent the estimated lottery premium for each year. The mean value of the lottery premiums equals -1.8%.

B5 Robustness analysis

To verify and validate our results, we repeat the Fama-MacBeth-regressions for different subsamples. We also include a regression model in which all variables, except for the *AlternativeLotteryDummy*, are standardized (mean equal to zero and standard deviation of 1). This modification is done to ensure that all variables are measured at equal scales, mitigating any potential biases or spurious results of variables contributing differently to the analysis. The four models are reported in Table B5.1 below.

Model (1) is the full sample regression model, while model (2) and (3) yield the subsample results measured over the periods of 1999-2010 and 2011-2020, respectively. Model (4) is the standardized model regressed on the full sample period.

Table B5.1: Robustness analysis of lottery stock premiums

This table reports the Fama and MacBeth (1973) monthly cross-sectional regression estimates for the value-weighted excess return on the alternative lottery-type portfolio, using different sample and variable specifications. Model (1) is regressed on the lottery portfolio across the full sample period, while model (2) and (3) are estimated over the periods of 1999-2010 and 2011-2020, respectively. Model (4) is specified by standardizing all input variables. Contemporaneous factor betas for the Fama-French risk factors (mkt, smb, hml), size (log of market value), book-to-market and lagged 6-month return are included as controls. T-statistics are reported in parentheses.

	Dependent variable:			
	Value-weighted excess return			
	(1)	(2)	(3)	(4)
AlternativeLotteryDummy	−0.501***	−0.566***	−0.283***	−0.038***
	(−5.386)	(−4.083)	(−2.653)	(−5.386)
mkt	0.937***	0.669***	0.888***	0.389***
	(5.008)	(4.521)	(2.724)	(5.008)
smb	−0.309	0.947**	0.785	−0.069
	(−0.874)	(2.478)	(1.622)	(−0.874)
hml	0.977***	0.609***	0.975***	0.267***
	(7.584)	(5.914)	(4.706)	(7.584)
Log_mktcap	2.133***	2.963***	3.478***	0.346***
	(16.333)	(13.197)	(15.658)	(16.333)
bm	0.883***	0.984***	1.005***	0.094***
	(11.852)	(8.109)	(10.784)	(11.852)
LaggedR	−0.011**	−0.015***	−0.022***	−0.011**
	(−2.500)	(−2.586)	(−3.534)	(−2.500)
Constant	13.465***	11.416***	10.689***	1.029***
	(6.207)	(5.761)	(3.533)	(6.207)
Model	Fama-MacBeth	Fama-MacBeth	Fama-MacBeth	Fama-MacBeth
Sample	Full	1999-2010	2011-2020	Full, standardized
Observations	671,039	380,025	291,014	671,039
R^2	0.230	0.223	0.266	0.230
Note:				*p<0.1; **p<0.05; ***p<0.01

When altering the sample period or standardizing input variables, all estimated coefficient signs stay consistent and all significance levels persist – except for the *smb*-factor. This result is explained in greater detail in section 6.2. By conducting separate regressions across the different subsamples, we confirm that the existence of lottery premiums is consistently significant over time. Moreover, as we interpret the same conclusion from the standardized regression, we are confident that our initial regression model is implemented without any measurement bias, and that our input variables contribute equally to the analysis.

Appendix 3: Regression Discontinuity Design

C1 Model specification of RD design

Table C1.1: Step-by-step specification of Regression Discontinuity model

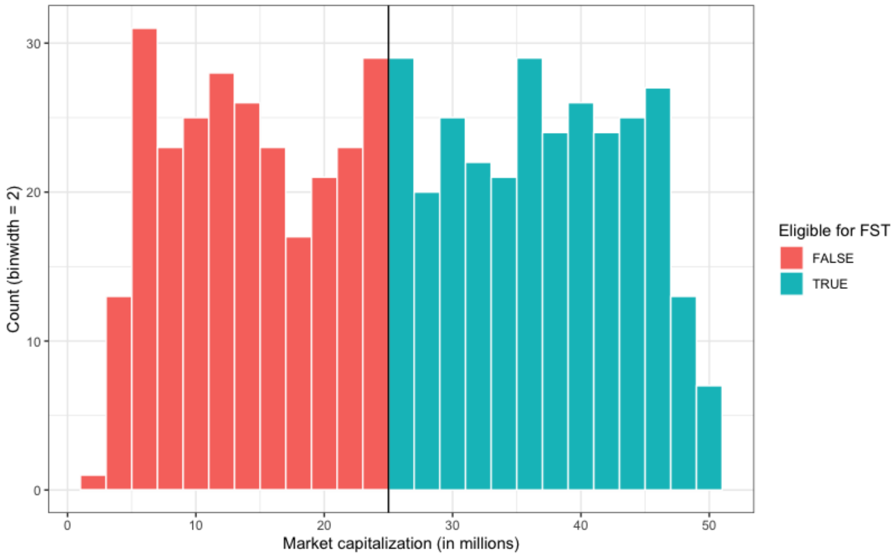
This table presents the step-by-step approach to specify the most appropriate regression model for our RD design. All models report linear regression estimates for the log of number of Robinhood investors holding stocks, proxying for retail investors' behaviour. Models (1) to (4) are regressed using different input control variables, one at a time. Model (4) yields the final model, in which an eligibility dummy indicating stock tradability, market capitalization, past maximum return measured over the previous month, stock price, idiosyncratic skewness observed over the last month are included as the covariates. Market capitalization is centred to indicate the distance of each observation from the eligibility cut-point of $25 million in market capitalization. T-statistics are reported in parentheses.

	Dependent variable:			
	log(Users holding)			
	(1)	(2)	(3)	(4)
EligibleDummy	-0.488^{**}	-0.385^{*}	-0.393^{*}	-0.368^{*}
	(-2.004)	(-1.802)	(-1.864)	(-1.761)
mktcap_centered	0.007	0.014^{*}	0.016^{**}	0.016^{**}
	(0.760)	(1.747)	(2.054)	(2.068)
Price		-0.107^{***}	-0.104^{***}	-0.104^{***}
		(-12.849)	(-12.604)	(-12.725)
MAX			0.014^{***}	0.013^{***}
			(3.925)	(3.654)
lagged_iskew				0.209^{***}
				(3.444)
Constant	5.858^{***}	6.317^{***}	6.108^{***}	5.886^{***}
	(43.685)	(51.374)	(46.119)	(40.250)
Model	OLS	OLS	OLS	OLS
Observations	552	552	552	552
Adjusted R^2	0.010	0.238	0.258	0.272

Note: *p<0.1; **p<0.05; ***p<0.01

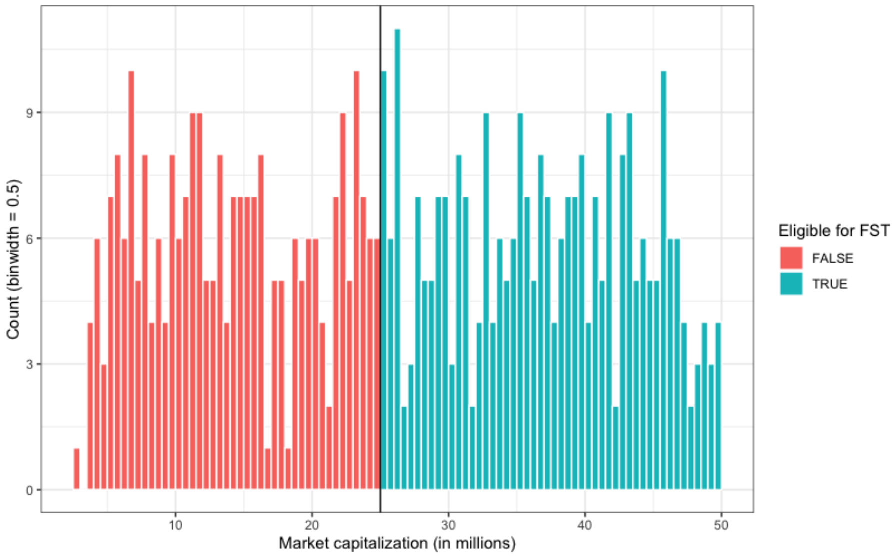
C2 Informal tests of optimal bin width

Figure C2.1: RD Running variable plotted with bin width equal to 2



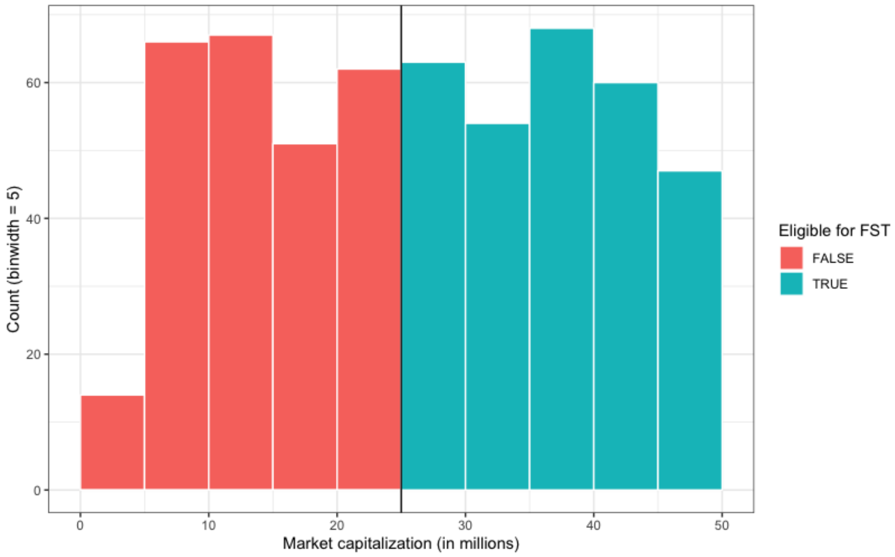
This figure shows the counted frequency of market capitalization values above and below the cut-point of $25M, represented by the black vertical line, using a bin width equal to 2.

Figure C2.2: RD Running variable plotted with bin width equal to 0.5



This figure shows the counted frequency of market capitalization values above and below the cut-point of $25M, represented by the black vertical line, using a bin width equal to 0.5.

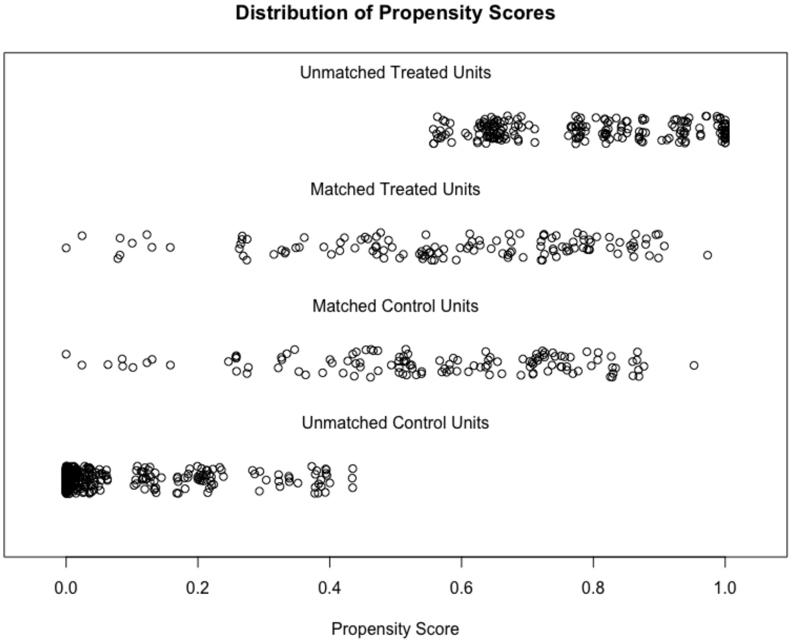
Figure C2.3: RD Running variable plotted with bin width equal to 5



This figure shows the counted frequency of market capitalization values above and below the cut-point of $25M, represented by the black vertical line, using a bin width equal to 5.

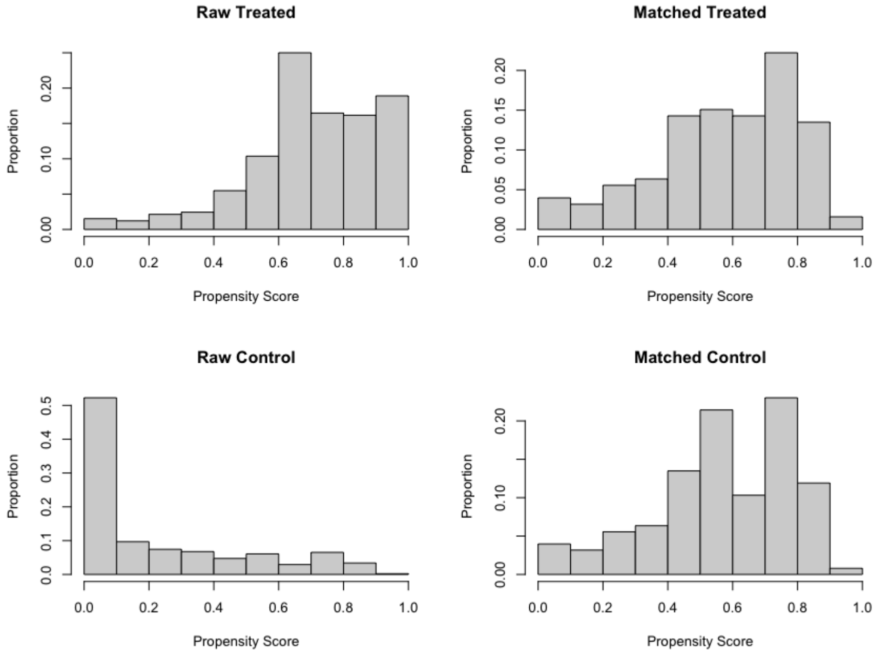
C3 Propensity score matching in RD design

Figure C3.1: Distribution of propensity scores for matched and unmatched samples in the RD design



This figure shows the distribution of propensity scores for both matched and unmatched treated and control units.

Figure C3.2: Histograms for matched and unmatched samples in the RD design

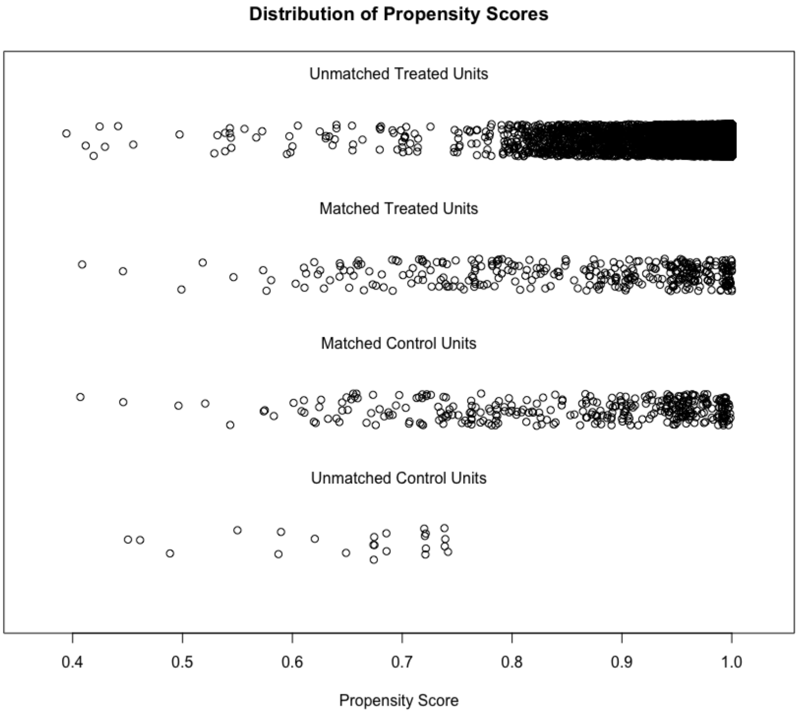


This figure shows the histograms for both matched and unmatched treated and control units.

Appendix 4: Difference-in-Difference

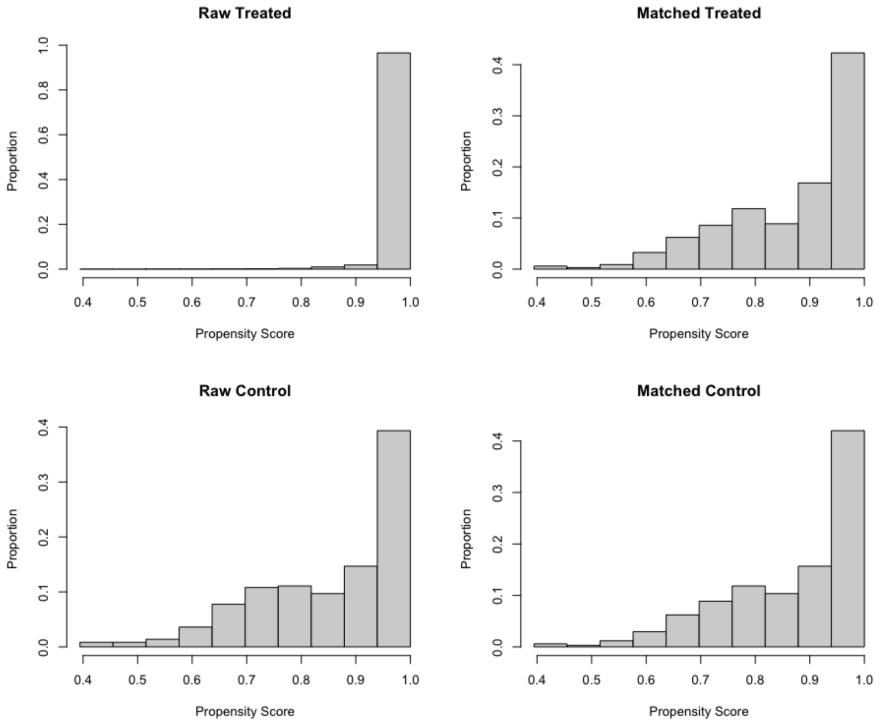
D1 Propensity score matching in DiD design

Figure D1.1: Distribution of propensity scores for matched and unmatched samples in the DiD design



This figure shows the distribution of propensity scores for both matched and unmatched treated and control units.

Figure D1.2: Histograms for matched and unmatched samples in the DiD design



This figure shows the histograms for both matched and unmatched treated and control units.