Incentive-Driven Trading: Behavioral Impacts in Financial Markets

Arina Wischnewsky[*]

Trier University and University of Luxembourg

Piotr Gański

University of Zurich

Thorsten Hens

University of Zurich and Norwegian School of Economics

Christoph Hölscher

ETH Zurich

Sandra Andraszewicz

ETH Zurich

October 8, 2024

Abstract

Individual risk attitudes vary based on numerous factors, such as the perception of the win-to-loss ratio, current financial situation, and cognitive biases. This study aims to investigate risk-taking and trading behavior under different compensation schemes using a dynamic experimental setting. Utilizing the Zurich Trading Simulator, a gamified trading tool, it is shown that compensation schemes significantly affect trading behavior.

[*] Corresponding author. Email: wischnewsky@uni-trier.de.

1 Introduction

Individuals vary in their investment behavior, based on their perception of the win-to-loss ratio, current financial situation, and the stakes size (Rabin, 2000). Personality defined within the framework of Big Five taxonomy also has an impact on both short and long-term investment (Mayfield et al., 2008). This leads to a situation in which people facing an identical expected value of an outcome vary in their decision-making. The impact of one's behavior and biases on utility function is part of a broader study, defined as Behavioral Economics (Ritter, 2003). The author proposes the idea that people are not always rational decision-makers when it comes to money. This irrationality can be observed in the finance industry when different stakeholders have different goals. A threat arises when the portfolio manager's objectives are not aligned with the investor's goals. Professional investors might compare performance with their peers, which leads to increased risk-taking and trading activity (Andraszewicz et al., 2022).

A misalignment occurs when a manager is not financially committed to his fund. Previous research shows that in such a situation he is likely to take on more risk (Ackermann et al., 1999). There is substantial literature supporting the thesis that managerial incentives influence risk-taking behavior, and that performance-related components of the salary encourage fund managers to take excessive risk (Kouwenberg & Ziemba, 2007). Furthermore, evidence points to the fact that incentive schemes with a threshold pose a moral hazard in the hedge fund industry. This is because the managers maximize their expected value of fees (Panageas & Westerfield, 2009). Furthermore, fixed-wage compensations, not aligned with payoffs, encourage minimal effort (Lazear, 2000). Fixed-wage compensation rarely produces positive incentive effects, since effort is not adequately rewarded (Bonner et al., 2000; Flannery & Roberts, 2021).

Measuring the activity of stock trading is important when assessing individual trading behavior. In a 2011 experiment, Jacobs and Weber (2012) focused on measuring the local bias of investors for firms trading on the German stock market. In their methodology, they determined trading activity by the number of buys and sells each day and compared the turnover between groups to look for a local bias. The authors concluded that investors in the German stock market found that local bias significantly impacts trading volume. However, their analysis was lacking in other ways of measuring trading activity. It did not mention individual trading frequency or the relation to risk. Those expansions, combined with risk analysis, could help better explain trading behavior. Conclusions based on a more rounded approach (trading behavior rather than only trading activity) will facilitate making more general predictions, scoping beyond the experimental sample.

The timing of the investment decision (buy/sell) is another component of trading behavior. There is a strong link between investor behavior and the short-term price movements of a stock. This has been shown by observing investment behavior right before and after the earnings announcement (Frieder, 2004). There is no real information that makes the price go up or down at the time of the decision. It is simply investors recognizing that others have a cognitive preconception of what the price will do irrespective of the fair valuation. They bet on other investors perceiving the earnings announcement as positive or negative, rather than based on fundamental analysis. It is another example of individual characteristics impacting trading behavior.

Another factor impacting the risk-taking aspect of trading behavior is the compensation scheme size and structure. Managers look at their decisions through the lens of their incentive scheme. This means that they perceive it as an investment problem, which they then try to solve (Carpenter, 2000). In situations where the firm's leadership compensation structure consists of options, its convexity plays an important role. It makes managers seek payoffs that are "away from the money," and thus lead to an overall risk burden increase. However, the paper has also shown that giving the manager more options makes him seek less risk (Carpenter, 2000).

This is in line with the theory that being heavily invested in the firm correlates with a more risk-averse approach to financial decision-making.

One of the most renowned and commonly referenced ways of measuring risk aversion is using multiple price listings (Holt & Laury, 2002). In such a task, subjects are asked to choose between lottery A or B. The choices vary in size and probability of a payoff. There is a calculation of the expected value of the lotteries, and often subjects choose a less risky one with a smaller EV, indicating risk aversion. The Holt and Laury paper serves as a great tool for identifying cognitive biases. If at price A, some individuals choose to sell, hold, or buy the stock and made different decisions, then it could be evidence of some cognitive bias. However, in a dynamic environment, this method is not the most applicable and thus will not be used in this experiment. The individuals will have to make hundreds of small decisions, which cannot be summarized as a choice difference between decision A and B. The methodology section explains in detail the alternative risk aversion measurement method.

When eliciting risk, there exists a trade-off between predictive power and noise based on the complexity of the method (Dave et al., 2010). This occurs when a trading task gets mathematically complex, and some participants might not fully understand it, thus creating noisy results. In the case of this experiment, the participants are from the fields of finance and economics, so their calculation skills should be adequate. Consequently, the manipulation variable in the experiment (incentive scheme) will have a somewhat complex structure. The reason for this is because the highly expected mathematical skills should help with understanding the experiment and thus limit the noise. Another aspect of eliciting risk preferences is the relationship with time (Andersen et al., 2008). The authors of the above article created a series of choice experiments and asked participants to make choices regarding the payoffs in their timing. The goal was to determine how individuals perceive present and future payoffs. One of the conclusions was that the preferences are not consistent. There are differences between short- and long-term preferences, as well as differences when it comes to the size of the reward. This hints that risk preferences revealed in this experiment might not be consistent with individual risk preferences in other settings.

Another aspect of decision making in experimental settings is risk aversion and expected utility theory (Rabin, 2000). Risk aversion refers to a preference for avoiding risk, especially in the case of potential losses. Expected utility is represented by a model explaining subjective utility expectations coming from potential outcomes. Both phenomena will potentially be observed in the following experiment. On one hand, in the pre-study, individual risk preferences will be measured, and given a large sample, some individuals should show high risk aversion. This means that, regardless of the experimental treatment, they will prefer safer outcomes. On the other hand, everyone will have their utility curve. This implies that the financial outcomes of the experiment are not the only sources of happiness for the participants. This will thus be taken into account in terms of this paper's limitations.

One of the most influential theories regarding the influence of price patterns on behavior is the Prospect Theory by Tversky and Kahneman (1974). The theory introduces the idea that people are more sensitive to losses than gains. Presentencing or "framing" can impact people's choices (Tversky & Kahneman, 1981). Loss aversion, which refers to the fact that people tend to feel more strongly about losses than they do about gains of the same size, can be observed in many situations. It implies that in a compensation scheme with a threshold, people will behave differently below and above it. For instance, if the reward for reaching a threshold is 5 CHF, then reaching it means a lot of additional utility. However, losing the "earned" 5 CHF would cause people to suffer a larger utility loss compared to the gain from the same value. This indicates that individual risk preferences are not transitive throughout the whole price pattern. Other drivers for behavior under uncertainty are representative heuristics (Tversky & Kahneman, 1974). It is the tendency to expect certain outcomes with a probability based on how similar it is to a stereotype. In the case of risk behavior in an experimental setting, it will be affected by participation in other experiments with risk measurement, as well as individual expectations of price development. Despite the price being fully randomized, participants will have a preexisting bias towards overestimating the likelihood of their expected price development.

Despite random sequences, some participants will interpret them as patterns. To minimize the probability of selection bias, the experimental stock price was randomized. This also ensures that the prices in the experiment are representative of a larger sample and not influenced by the experiment designer's bias (Thaler & Sunstein, 2003). Furthermore, the generalizability of the findings can be used to draw conclusions that extend beyond just the sample population.

The above serves as a starting point for the research question. However, experiments based on historical data and primary lab research are not the same. To answer the research question (discussed in more detail in the Design section): To what extent do incentive schemes impact trading behavior? one more publication needs attention. It serves as a leeway between the theoretical framework and already existing experimental evidence. The article creates an overview of already existing literature on the relationship between monetary incentives and effort/task performance. In addition to the literature review on the topic, the authors suggest that the meta-analysis provides mixed results (Bonner et al., 2000). This means that the methodology for comparing risk and incentive schemes should be individually tailored to the experiment. They develop a framework that should guide experiment designers. In the framework, monetary incentives impact effort, which in turn impacts task performance. However, many other variables affect both perception of monetary incentives and task performance. These variables include person variables, task variables, environmental variables, and incentive scheme variables. To fully understand the risk-reward relationship, all the above-mentioned factors should be taken into consideration. In my experiment, effort will be a proxy for trading activity, since that is the only action participants will be able to take. However, as explained in the methodology section, effort should not be judged identically for all participants, since each compensation scheme has its own goal based on the payoff scheme. Moreover, since risk preferences share the structure of major psychological traits, individual characteristics will play a big role in this experiment (Frey et al., 2017). One such example is that risk preferences have a similar psychometric structure to intelligence. However, Frey emphasizes the risk measurement method. In other words, understanding individuals' risk preferences and being able to affect them can have implications beyond Behavioral Economics.

1.1 Pilot Study

Additionally, to the previously mentioned literature, the experiments from a Master's student conducting prior research with the Zurich Trading Simulator will be used as a point of reference for setting up the incentive schemes (Dousolier, 2021). The primary finding suggests that risk-taking varies based on participant compensation. More specifically, in contrast to people working as fund managers trading their fund's money, individuals trading with their "own" money seem to be more risk-averse. Individuals whose compensation is based on the threshold they must reach to get a bonus seem to take on more risk than the average across participants, especially if the threshold is difficult to achieve (i.e., a high threshold). Moreover, they trade more frequently and with higher volume, which—given the existence of transaction fees—can be suboptimal.

The issue with the above research is that threshold manipulation was not set up optimally. In the experiment, the bonus was set up as a non-monetary value: it was a credit incentive for students, which they required to obtain their degree at the University of Basel. This can be seen as a limitation, since non-monetary and monetary incentives do not necessarily work in the same way.

However, overall, the results are in line with literature concerning risk-taking for different types of hedge funds. Incentive fees reduce managers' implicit level of loss aversion. On the other hand, if a manager's own stake in the fund is more than 30%, risk-taking is reduced (Kouwenberg & Ziemba, 2007). The last conclusion from the pilot studies is that there seems to be no dominant strategy for the control and treatment groups. No group seemed to choose a trading strategy that would provide a statistically significant higher return on investment than the average across all participants. The pilot results have been promising; however, the incentive schemes selected have not been optimal for reflecting real-life behavior. The reason for this is that only the linear compensation mirrors a potential scenario for an individual investor.

1.2 Research Question and Hypotheses

The literature review, combined with the pilot study, led to the development of the following research question:

To what extent do incentive schemes impact trading behavior?

The following hypotheses will help answer the research question above:

Hypothesis 1a – The more aligned a compensation scheme, the fewer the number of transactions.

Hypothesis 1b – The more aligned a compensation scheme, the smaller the average size of transactions.

Hypothesis 1c – The more aligned a compensation scheme, the smaller volume of shares traded.

Hypothesis 2a – There is a negative effect of aligning the incentive schemes on risk.

Hypothesis 2b – There exists a strong correlation between self-reported risk and in-game risk.

Hypothesis 3a – In compensation with a fixed bonus, individuals will suspend all risk-taking and trading activity after reaching the threshold.

> **Hypothesis 3b** – Participants in experimental conditions with a threshold will behave more risk-seeking under the threshold and less risk-seeking just above the threshold.

Hypotheses 1a, 1b, and 1c are based largely on previous findings indicating that misaligning incentive schemes leads to higher risk-taking (Ackermann et al., 1999). Moreover, being disconnected from the monetary outcome does not promote risk aversion (Holt & Laury, 2002). The risk-taking in this experiment can be partially observed via trading activity. Trading more often, with higher frequency, and with higher volume implies higher risk-taking behavior.

Hypothesis 2a is based on the same principle but focuses on looking at risk as a share of the risky asset of the whole portfolio, rather than through trading activity. The hypothesized effect on risk has an additional noise component created by individual biases of the expected future price movement (Frieder, 2004). **Hypothesis 2b** adds the component of self-reported risk. Questionnaires are a common way to measure risk (Charness et al., 2013). If individuals answer honestly, their self-reported risk and in-game risk should be correlated.

Hypotheses 3a and **3b** focus on common compensation types, namely performance-based compensation. On one hand, fixed compensations invite minimal effort (Lazear, 2000). In the Zurich Trading Simulator setting, this indicates that when reaching the performance threshold, participants should no longer be interested in making any more trades or putting in any more effort. This is the expected outcome for **Hypothesis 3a**. If participants in the experiment behave like real-world managers, they will look at the structure of their compensation as an investment problem (Carpenter, 2000). They will adjust their behavior according to their individual solution to that problem. Threshold conditions by design incentivize reaching the threshold; thus, participants will aggressively try to reach it and then limit their exposure to stay above it. This is in line with Prospect Theory by Tversky and Kahneman (1979), which states that people have different utility curves depending on whether they have a positive or negative monetary outcome. The expected outcome in **Hypothesis 3b** is higher risk-taking behavior below the threshold and lower risk-taking behavior above the threshold.

2 Design and Method

2.1 Design

2.1.1 Experiment

The goal of this study is to replicate the real-life trading behavior under various incentive schemes. This will allow us to measure the revealed risk preferences of the participants. The three types represent common compensation schemes found in the finance industry. In this experiment, there will be three compensation schemes (compensation 1 serving as the benchmark):

1. Linear compensation fees - replicating the behavior of individual investors (Linear)
2. Fixed wage + fixed bonus (2% threshold) replicating the compensation of professional investors (Fixed)
3. Fixed wage + linear bonus (2% threshold) replicating the behavior of professional investors in managerial positions (High-Watermark)

The ”Linear” scheme imitates individual investors using a trading platform with no transaction fees. The ”Fixed” and ”High-Watermark” groups represent other incentive scheme types commonly found in the finance industry. The comparison between the second and third schemes furthers the research, suggesting that aligning the goals of fund managers and investors leads to more optimal risk-taking behavior (Jensen, 1990). The risk preferences over trading days in the three conditions are illustrated in the figure below.

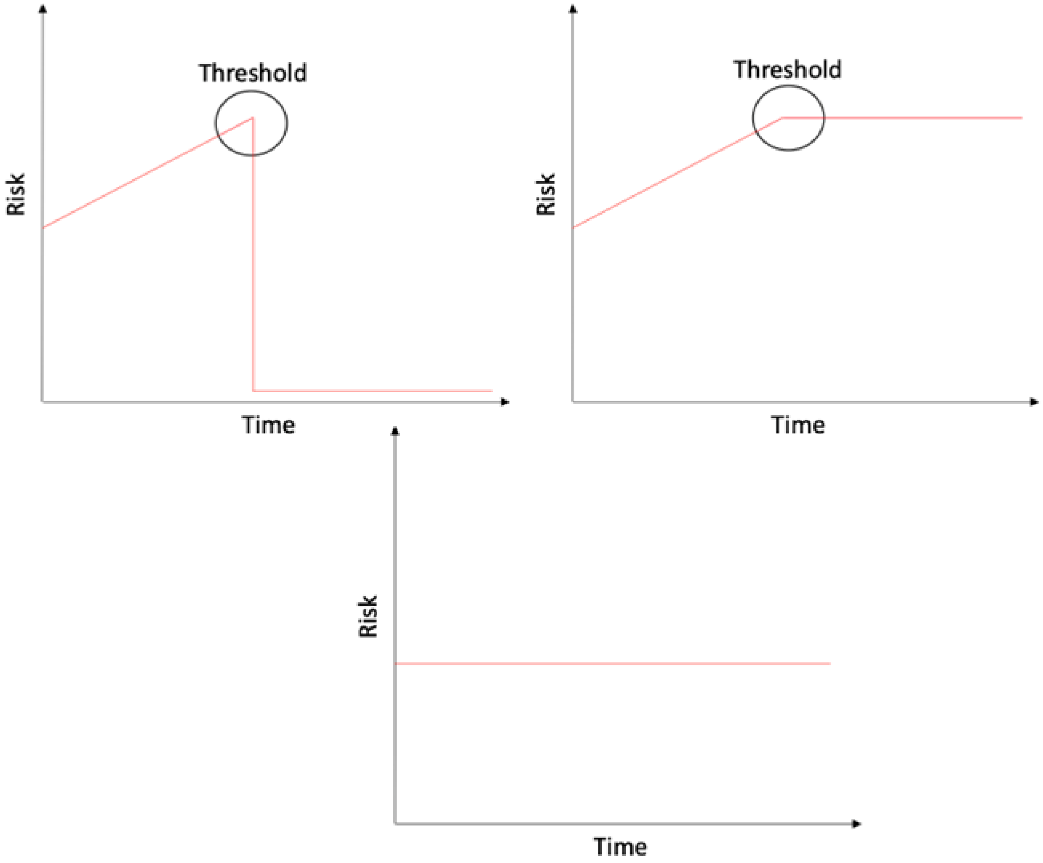

Figure 1: Risk Preference (risky asset/cash) over time (trading days) for various compensation schemes.

2.1.2 Pre-experiment Study

Before the main experiment, a pre-experiment study was conducted (Appendix 3). It had two main objectives: test how likely participants could reach the 2% threshold and determine how to compensate each group so that, on average, all could earn a similar amount of money. The results (N=22) showed that the mean return in the game is 2.44%, and the following payment schedule yields, on average, a similar result of 8 CHF for each incentive scheme:

Linear: Show-up fee = 6 CHF and 1 CHF for every 1% increase in portfolio value.

Fixed Bonus: Show-up fee = 5 CHF and 5 CHF for reaching the 2% threshold.

Watermark: Show-up fee = 7 CHF and 1 CHF for every 1% increase in portfolio value after reaching the 2% threshold. (Appendix 3)

2.1.3 Linear Compensation

The goal of this compensation scheme is to mimic the real-life circumstances of an individual trader. In this compensation scheme, the participant is fully responsible for their pay. After receiving the 6 CHF show-up fee (for participating in the experiment), all potential gains transfer directly into a monetary payoff. The optimal behavior depends on individual risk preferences explained by indifference curves in prospect theory. Prospect theory, developed by Tversky and Kahneman (1979), describes the hypothesized behavior of participants in this scheme. The risk preferences and loss aversion should be linear regardless of the profit achieved.

2.1.4 Fixed Bonus

The goal of this compensation scheme is to elicit more risky behavior compared to linear compensation, as seen in the pilot studies. Participants will now optimize their behavior based on the threshold according to de Figueiredo Jr et al. (2019). According to the literature, the distance to the threshold will now be a key determinant of risk-taking behavior. We expect participants to optimize their risk-taking to meet the 2% threshold. If participants behave fully rationally, they should buy/sell assets until reaching the threshold; after that, they should stop all transactions. When participants keep trading after reaching the threshold, it may indicate a lack of understanding of the trading task or irrational behavior.

2.1.5 High-Watermark Compensation

This compensation scheme reflects the real-life incentives of executives in the financial industry. In a typical hedge fund, there exists a 2/20 rule and a high-water mark. This means that for simply running the hedge fund, investors agree to pay

2% of their invested money as a management fee and 20% of all profits above a predetermined high-water mark. However, in a bearish market, investors do not pay the 20% fee if returns are below the high-water mark.

The incentives of investors and managers are then somewhat misaligned. While investors lose part of their invested share, managers still receive the management fee (fixed pay) without the bonus. In such scenarios, participants behave optimally if they trade until reaching the threshold value of 2%. Any decisions made after reaching the threshold depend on individual risk preferences, similar to the linear compensation scheme.

2.2 Participants

Comparing means with ANOVA and non-parametric t-tests requires a certain number of participants to obtain the desired effect. According to the power analysis, assuming the following parameters: effect size of d = .2 and a power of .95 (α = .05), the experiment required between 180 and 250 participants. However, since only financially literate participants were eligible for the study, the focus group consisted of finance and economics students. The final number of participants was 201, split into: 110 males, 90 females, and 1 non-binary person. The median age was 24 years old, with the youngest participant being 18 and the oldest 55. They were recruited through the mailing list of the Decision Science Laboratory at ETH. The only condition was that they had to be students with some preexisting financial knowledge (e.g., Master of Management and Economics at UZH). The participants took part in the experiment, which was approved by the Ethics Committee of ETH (Appendix 1).

2.3 Materials

2.3.1 Zurich Trading Simulator (ZTS)

ZTS is a dynamic trading tool developed by the Chair of Cognitive Science at ETH Zurich for simulating the stock exchange. Its objective is to replicate a real-world equity trading experience in an experimental setting. Subjects are presented with the possibility to trade at different levels of volume. Moreover, they are not limited by trading only once per experimental day (Andraszewicz et al., 2023). The primary reason for using a dynamic tool, instead of a simple equity allocation choice, is that a simulated experience improves risk perception, allowing for the measurement of trading activity apart from risk-taking (Hogarth et al., 2015). Participants will be presented with a ZTS-generated single stock market and asked to decide when to buy or sell as the price keeps evolving. The goal will be to measure their risk-taking and trading activity. Further analysis in the Results section will focus on whether those decisions vary based on the compensation scheme.

2.3.2 Price Path

The price path in this experiment is determined by a random walk model. The reason for having a random price path, rather than a predetermined one, is to exclude any bias caused by participants being familiar with it. However, this does not mean that the price path is unrealistic. The Random-Walk Theory states that a market price is independent of any previous market-price patterns (Malkiel, 1999). This has been empirically tested (Van Horne & Parker, 1967), showing that a trader using technical analysis to predict future trends based on past trends cannot achieve higher gains than a buy-and-hold investor. Another benefit of randomly generating a price path is that it can elicit certain behavior from participants. The experiment focuses on the impact of incentives on trading behavior, with incentives based on performance in trading tasks. Certain thresholds (explained in more detail in the compensation section) need to be attainable for most participants. Figure 2 below shows the selected price path and its descriptive statistics.

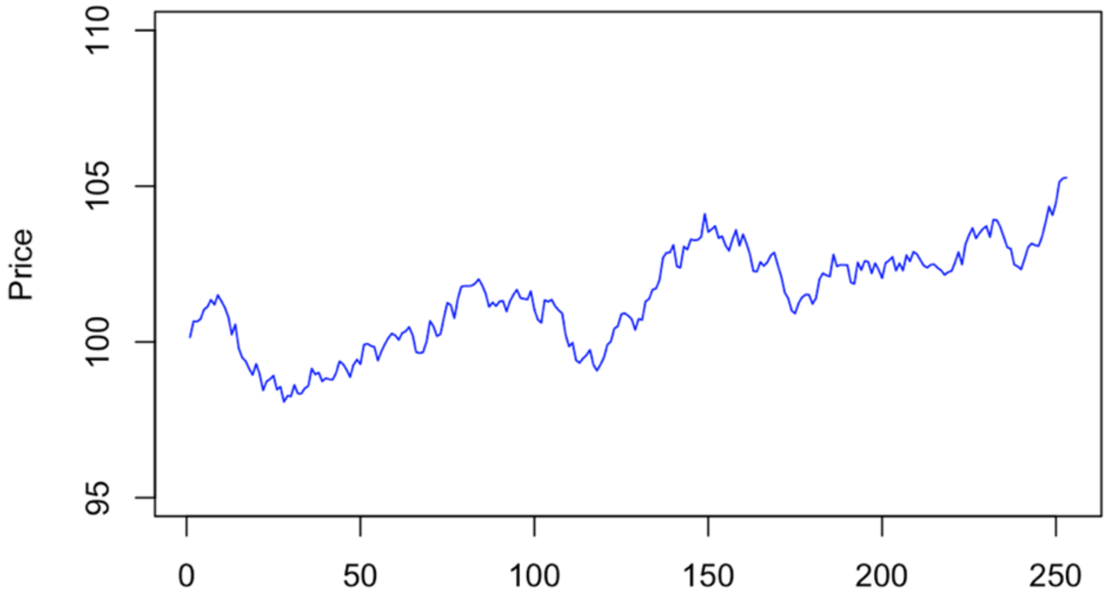

Figure 2: Price Path

	Minimum	Maximum	Mean	Standard Deviation
Price ($)	98.08	105.25	101.4	1.58

Table 1: Descriptive Statistics for the Price Path

Figure 1 shows that there's a very low risk of negative outcomes, which would be difficult to enforce in an experimental environment. All participants should be earning a similar amount of money, regardless of the incentive scheme, which in this case is around CHF8.

2.3.3 Risk Measurement

Risk elicitation was conducted in two separate ways. First, participants were asked to determine their risk preferences on a scale from 0 (risk-averse) to 100 (risk-loving) (Appendix 2). Questionnaires like this are a common method to measure self-reported risk propensity (Charness & Gneezy, 2012). Second, risk was measured in the experimental task using a methodology similar to the social comparison paper by **andraszewicz2022<empty citation>**. This method, developed by Charness and Gneezy (2012), defines risk as a share of risky assets compared to non-risky assets (cash) at a point in time. Most of the data and the Results section will focus on how compensation affects this experimental measure. All participants start with the same risk value of 0.5, corresponding to 50% of their portfolio invested in risky assets. The risk is thus affected by both price development and participant decision-making (Appendix 3.3).

2.3.4 Surveys

Before the ZTS experiment, participants completed a Qualtrics survey (Appendix 2). The survey's goal was to measure individual financial literacy and self-reported risk preferences. It also explained the procedure and conducted a manipulation check. The post-survey recorded individual characteristics such as age, gender, and education, and served as a platform for feedback.

2.4 Procedure

2.4.1 Experiment

After completing the financial knowledge survey, participants moved on to the experiment. They were randomly allocated to a compensation scheme and shown the instructions. In the ZTS tool, individuals chose how much of the risky asset they wanted to buy or sell. All participants were shown the same price movements in two separate rounds. Finally, rewards were calculated according to the incentive scheme and paid out together with the show-up fee.

2.4.2 Trading Activity

Trading activity is largely unexplored in the dynamic trading game literature. ZTS does not impose a hard cap on the frequency of trades, meaning the maximum trading frequency depends only on the participant's ability to process price developments and press the trading buttons. However, experimental limitations exist regarding trade size. Each trade can only be executed in volumes of 1, 5, or 20 shares. The last variable of interest is overall trade volume, which is part of a broader investigation into trading behavior and is closely tied to individual risk preferences and rational behavior.

3 Results

There are three hypotheses investigating whether the trading behavior (number of trades, volume, and transaction size) of individuals is dependent on their compensation scheme. I used the Shapiro test before choosing the appropriate hypothesis test. The significant results at 1% indicate the data is not normally distributed (Shapiro, 1964). Thus, non-parametric tests are most appropriate. The primary tool for comparison was the Wilcoxon test for mean and median comparison between groups. A Wilcoxon test was used to compare the median differences in risk behavior between the three experimental groups in the ZTS trading simulator (Wilcoxon, 1945). The Kruskal-Wallis test was used to compare whether the groups came from the same distribution (Kruskal & Wallis, 1952). Lastly, the Fligner-Killeen test was used to assess whether the experimental groups have significantly different variances (Fligner & Killeen, 1976).

Hypothesis 1a: Difference between the number of trades.

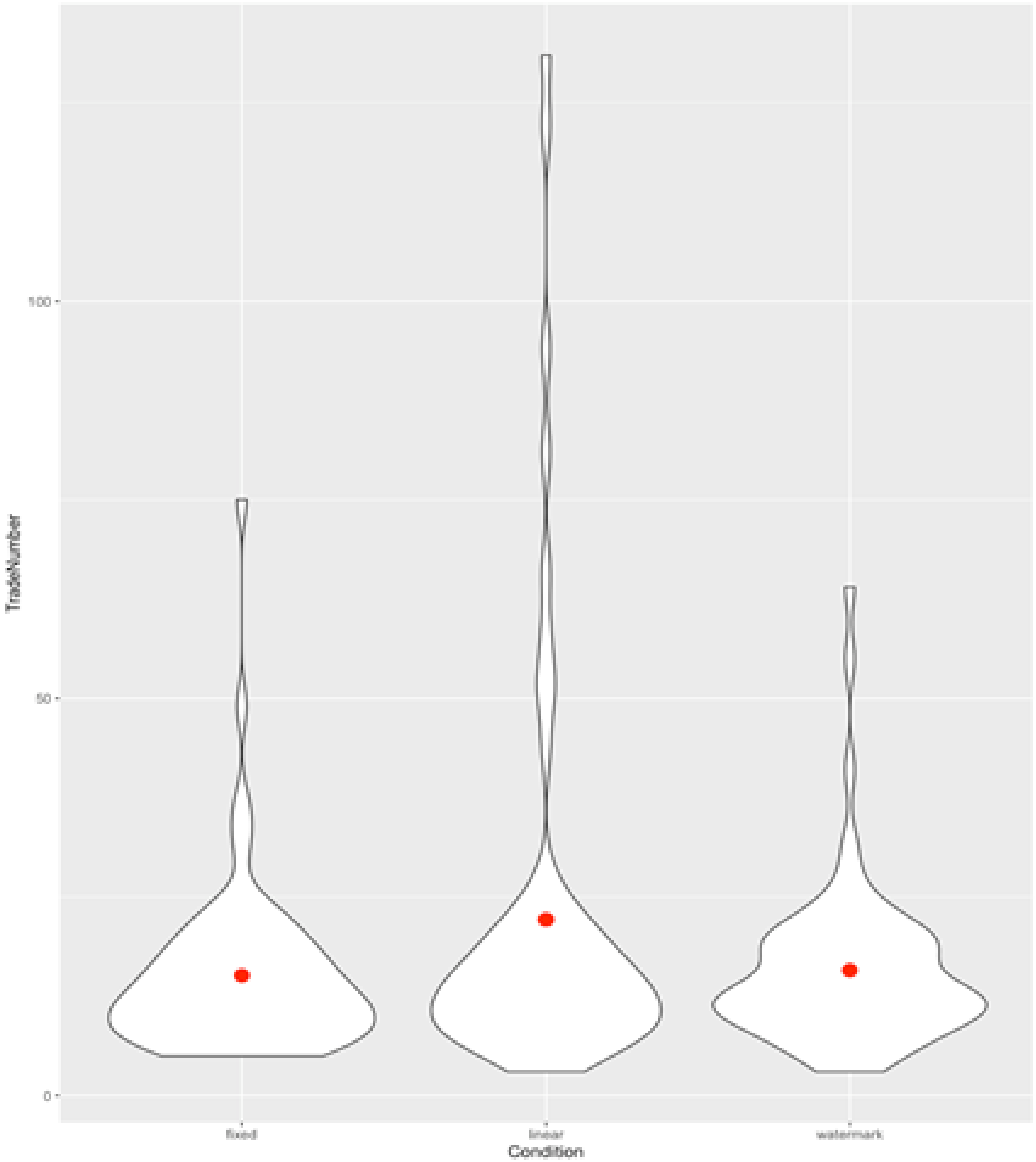

Figure 3: Difference in the Number of Trades between the three experimental conditions. (The red dot represents the mean number of trades)

In terms of hypothesis testing, the output of the Wilcoxon test in R indicates that there are no statistically significant differences between the three groups in terms of the mean number of transactions (p = 0.92). Moreover, the hypothesis testing for variance difference also shows no statistically significant results (p-value = 0.274). However, the mean value for the linear group seems to be higher. This is mostly due to a few outliers with more than 80 trades, which skew the mean upwards. This suggests that participants in the linear condition behave more precisely since their payoff is not dependent on a bonus and is thus not represented by a utility step function.

Hypothesis 1b: Average size of transactions.

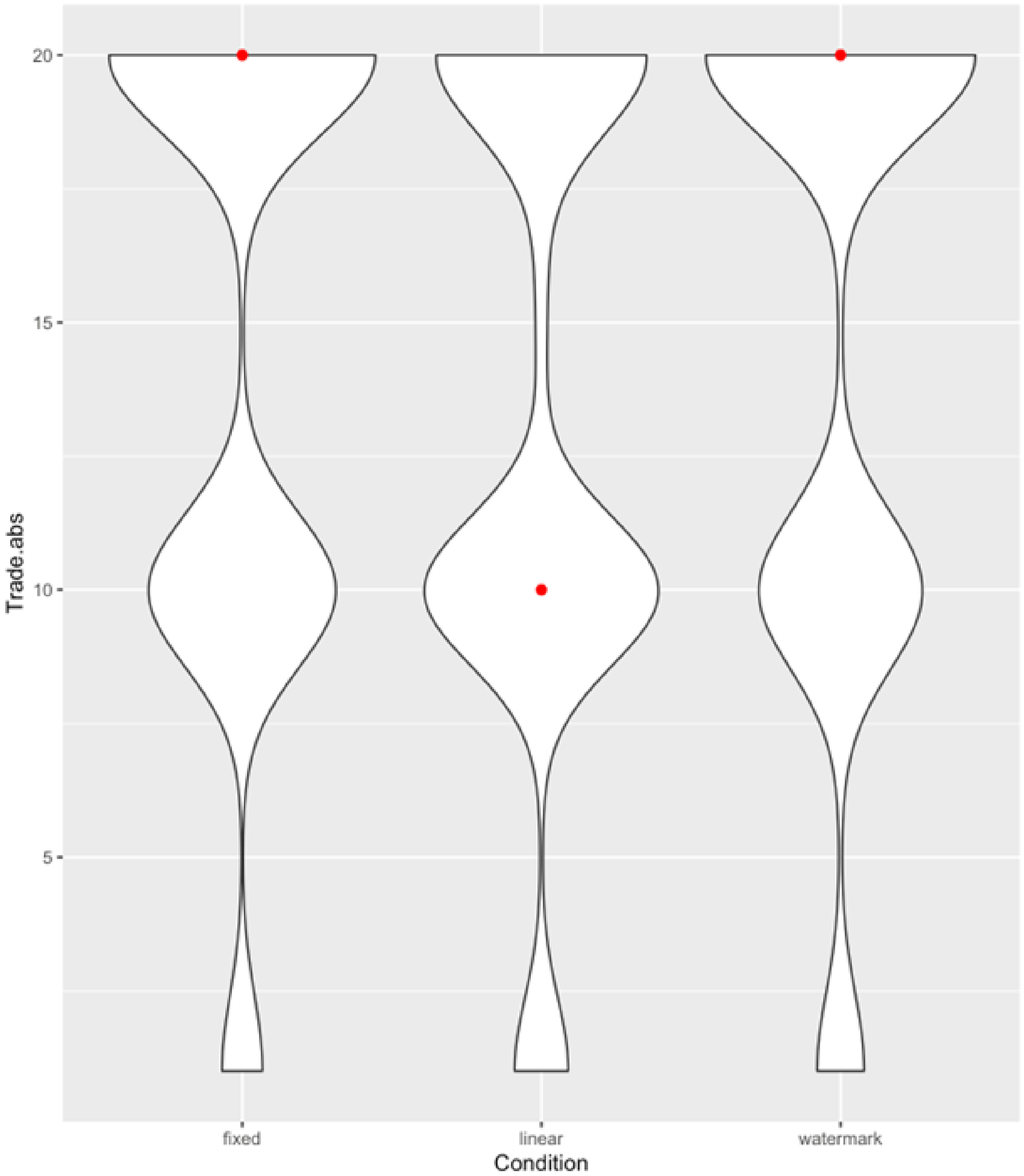

Figure 4: Average Size of Transactions by experimental condition. (The red dot represents the median trade button size)

This hypothesis investigates further whether linear participants are incentivized to be more precise compared to the other experimental conditions. Specifically, it relates to the size of the trading button they can press. Each participant can choose at any point in time to trade either 1, 10, or 20 stocks. The Wilcoxon test shows a statistically significant difference at 1% (p = 5.3e-08, p = 1.2e-08) between the linear and fixed and watermark groups. The difference between the watermark and fixed groups is not statistically significant (p = 0.76). The other two non-parametric tests are also significant at 1%. The Kruskal-Wallis test indicates that the experimental conditions have probabilistic differences that are not due to chance alone (p-value = 1.145e-10). The Fligner-Killeen test indicates statistically significant differences in variances (p-value = 6.797e-05). There are no differences between the watermark and fixed schemes.

Hypothesis 1c: Average Transaction Share Value.

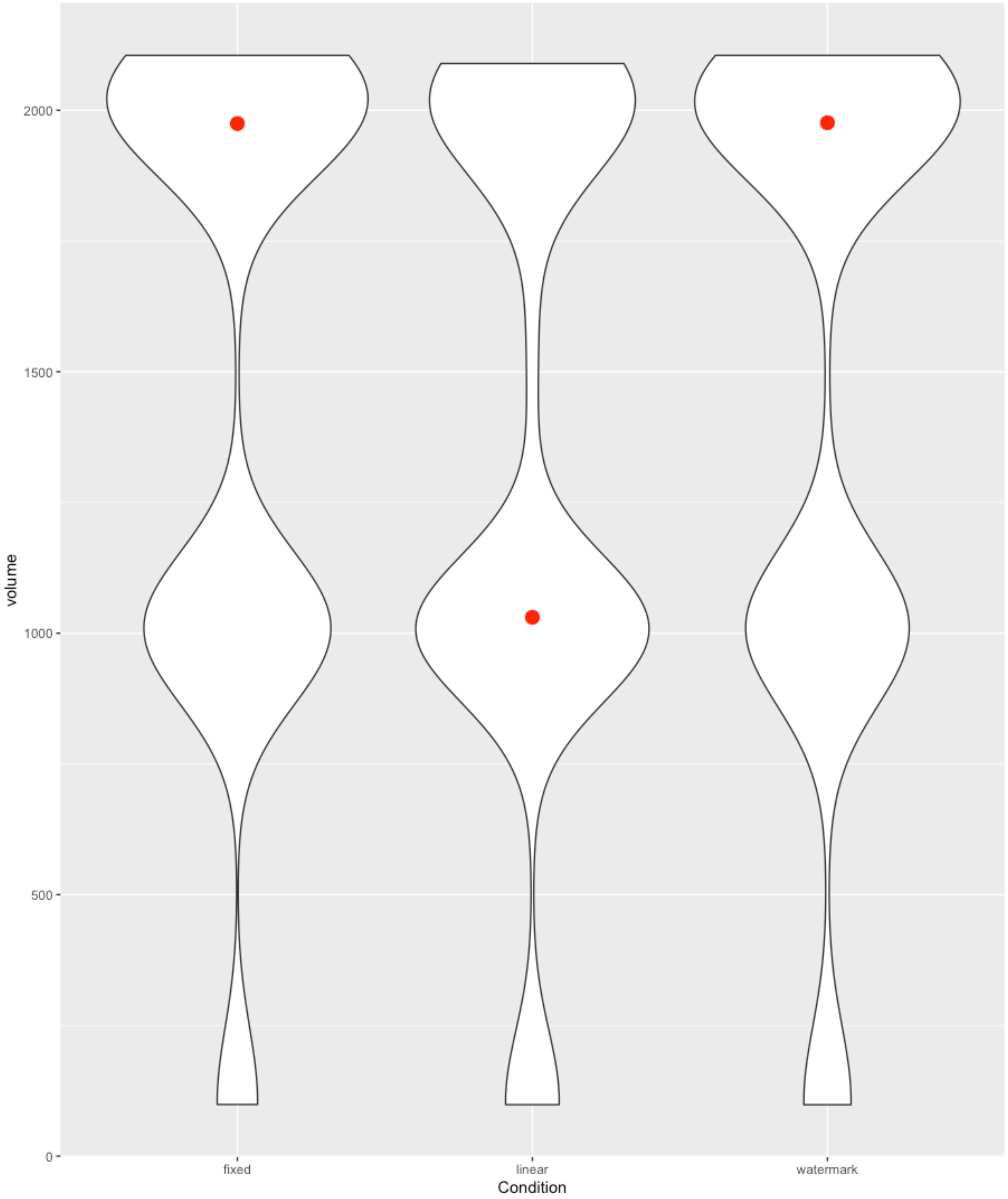

Figure 5: Average Transaction Share Value by experimental condition. (The red dot represents the median)

The variable volume is calculated by multiplying the share price by the size of the trade button. The resulting distribution is very similar to Hypothesis 1b. There are statistically significant differences between the linear compensation scheme and the other schemes at 1% (p = 0.002). The variance (p-value = 0.04656), distribution (p-value = 6.084e-08), and median between linear-fixed (p-value = 1.7e-06) and fixed-watermark (p-value = 3.1e-06) all vary statistically significantly at 1%. There are no differences between the watermark and fixed schemes.

Hypothesis 2: The focus of Hypothesis 2 is on the risk-taking aspect of trading behavior. Risk is represented as the share of assets vs cash in the portfolio. The number of participants for this hypothesis is 201, and the risk data is not normally distributed.

Hypothesis 2b: Strong correlation between self-reported risk and in-game risk.

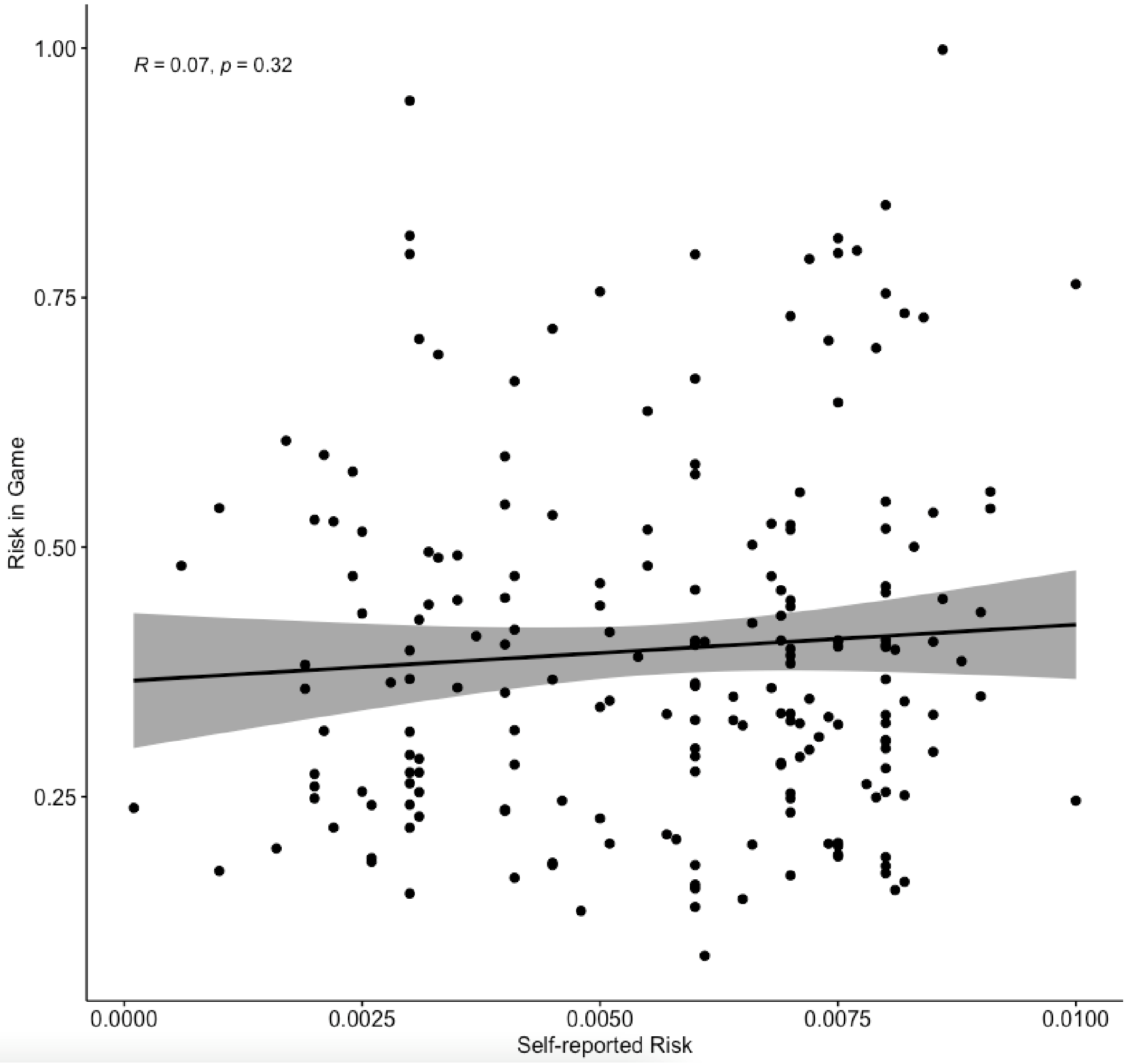

Figure 6: Correlation between Risk in Game and Self-reported Risk

Normality test Shapiro: p = 0.001, which indicates that risk is normally distributed. Similarly, self-reported risk is also normally distributed (p = 0.001). However, based on the graph, there is no significant correlation (p = 0.32), meaning the correlation could be equal to zero. This suggests that the self-reported risk does not reflect individuals' actual risk preferences accurately. This poses a threat in situations such as financial institutions using surveys to gauge risk preferences before clients commit to funds or pension plans. If there is no correlation between actual risk preferences and reported preferences, survey results may introduce significant bias.

Hypothesis 2a: Effect of aligning the incentive schemes on risk.

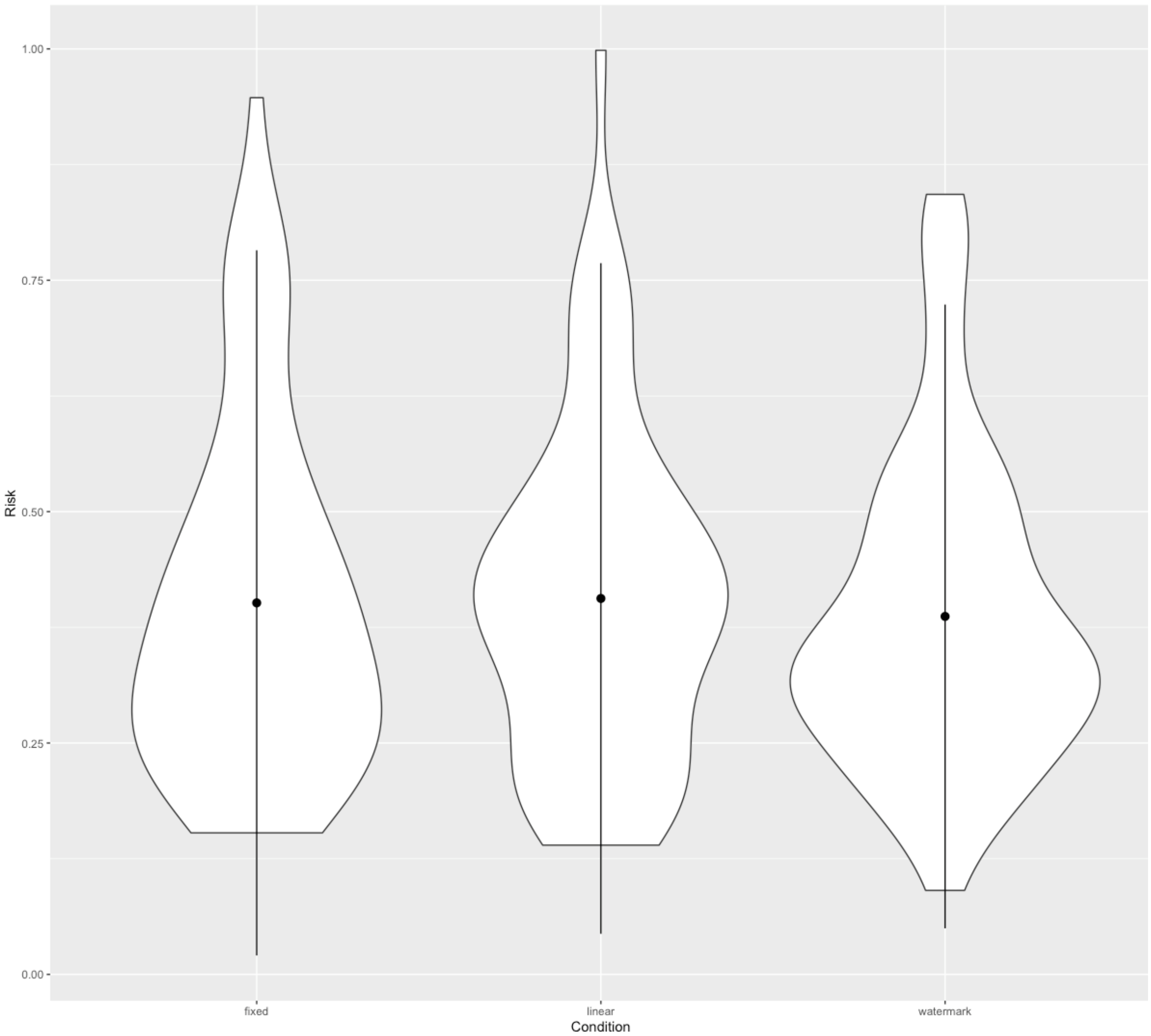

Figure 7: Mean differences between Aggregated Risk and Experimental Conditions. (Black dot shows the mean)

The hypothesis that the compensation scheme alone affects average in-game risk-taking behavior can be rejected. Averaging out the risk through the experiment shows no significant differences between distribution (p-value = 0.8042) and variances (p-value = 0.5779) across experimental conditions. The next step is to examine risk at individual points in time rather than as an average. The graph below shows risk during each trading day for the three conditions, as well as the price path.

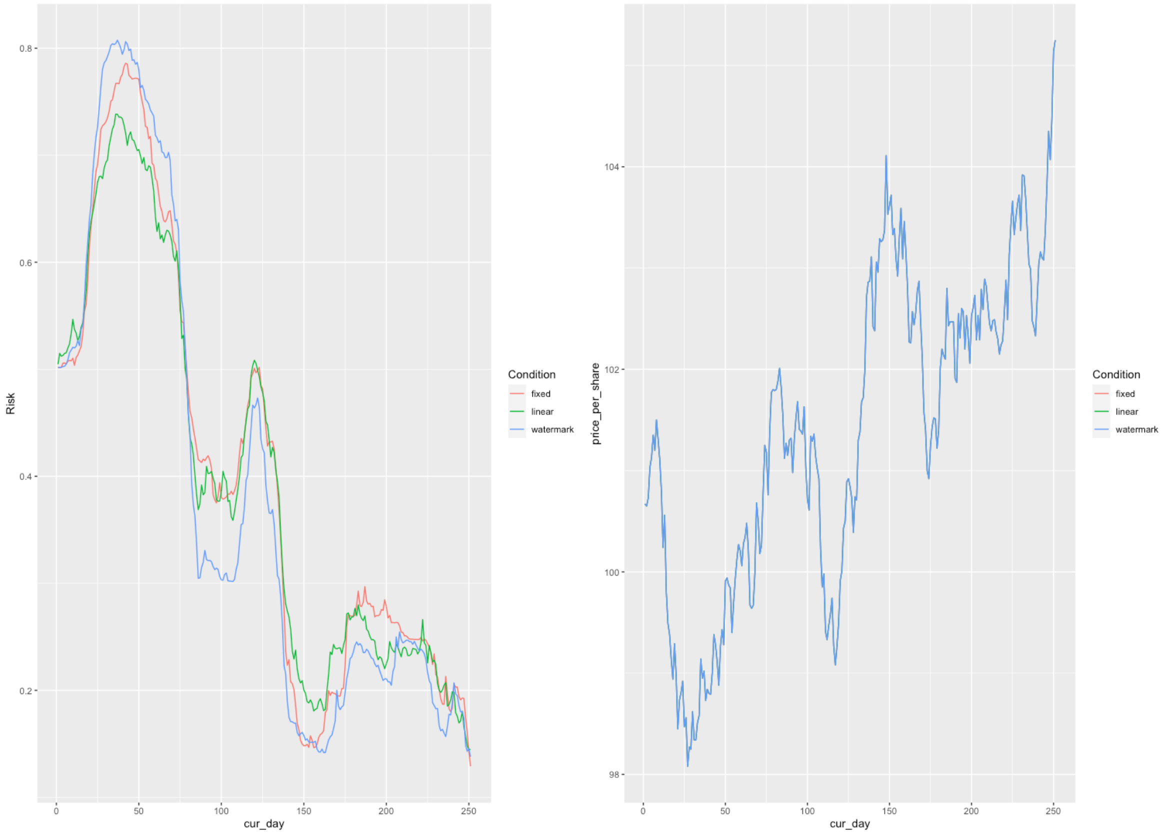

Figure 8: Risk over Days Traded, separated by Experimental Conditions

In terms of hypothesis testing, tests for mean (p = 0.14, p = 0.14, p = 0.65), variance (p-value = 0.2566), and distribution (p-value = 0.1441) differences are all insignificant at the 5% level. However, this does not necessarily mean that no differences exist. There are visual differences, especially between the watermark condition and the others. To identify actual differences, the sample needs to be split further.

Hypothesis 3b: Threshold Behavior.

The threshold was set at 2%, so the dataset was split into two components: behavior after reaching 10200 in-game currency (2% threshold) and behavior below it. As with other hypotheses, the number of participants for this test is N = 201. The table below summarizes aggregated risk below and above the threshold.

Risk	Condition	Threshold
0.39	Fixed	Below
0.40	Linear	Below
0.37	Watermark	Below
0.18	Fixed	Above
0.24	Linear	Above
0.22	Watermark	Above

Table 2: Risk Statistics Below and Above the Threshold

As seen in the table, all groups decreased their portfolio risk after reaching the threshold. The largest difference is for the fixed condition, where the opportunity cost is highest if the 2% threshold is not achieved (leading to 5 CHF less in post-game earnings). This indicates that the experimental treatments were effective, serving as a proof of successful manipulation check. Further analysis will focus on individual behavior both below and above the threshold.

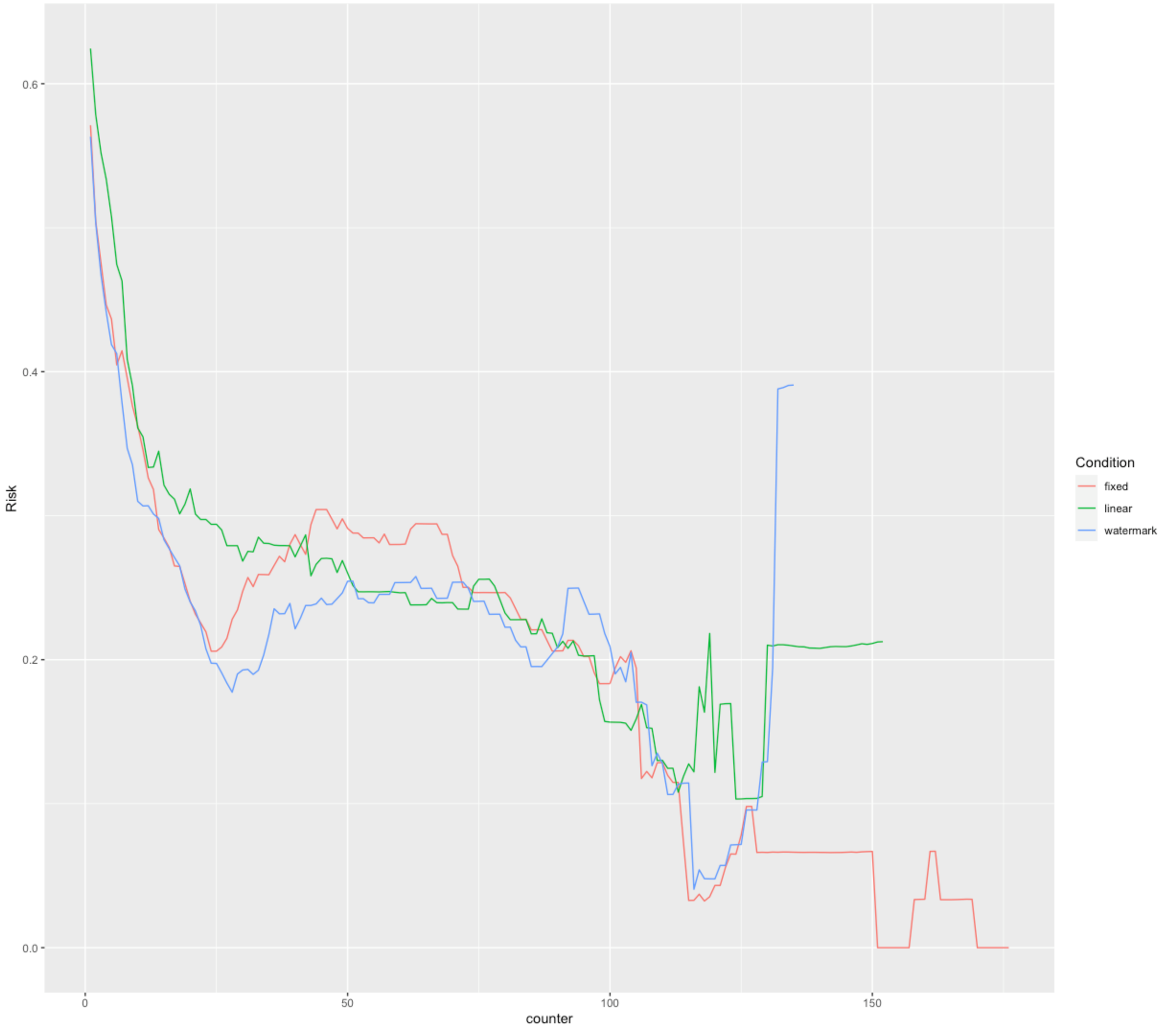

Figure 9: Risk over the Trading Days Above Threshold

The x-axis shows risk, while the y-axis shows days above threshold. The graph clearly shows that as the days over the threshold increase, the difference between groups becomes more visible. After around day 120, the three groups show the biggest differences. Watermark takes the most risk, and fixed takes the least risks above the threshold. This aligns with the fixed compensation optimal strategy, which is to suspend all trading after reaching the threshold. However, this strategy should be implemented immediately after reaching the 2%, starting from day 1. The trading activity highlights two key points:

1. Lag in cognitive response (participants are slow to recognize that they should suspend all trading and switch all assets to cash).

2. Irrational behavior (participants did not fully understand their optimal strategy in such a situation).

Due to this bias, the comparison should be made at the aggregated risk levels.

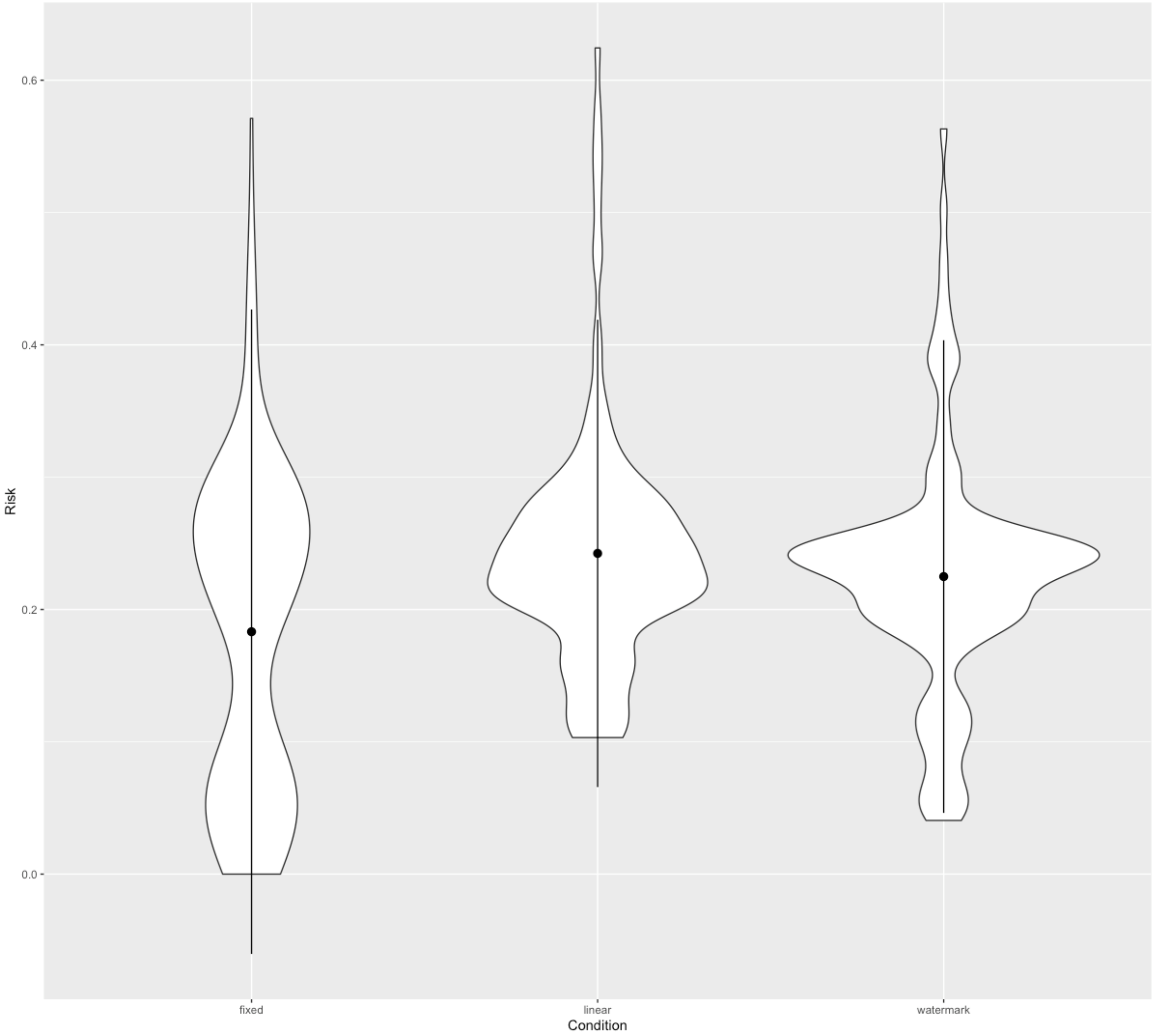

Figure 10: Risk by Condition Above the Threshold (Black dot indicates the mean risk).

The graphs visually confirm that the fixed compensation group takes the least risk above the threshold, again demonstrating that the manipulation in this experiment has worked. In terms of hypothesis testing, the hypothesis that the distribution (p-value = 0.00327) and variance (p-value = 0.002891) are different between groups cannot be rejected at 1%. The hypothesis for differences in means cannot be rejected at 1% for the differences between the fixed and linear schemes (p-value = 0.00087), at 10% for the differences between the fixed and watermark schemes (p-value = 0.05598), and at 10% for the linear and watermark schemes (p-value = 0.10202). The biggest differences are observed between the fixed and linear schemes, which is in line with the hypothesis, as participants in the linear scheme are not aware of the threshold, whereas participants in the fixed scheme have the highest opportunity cost associated with the threshold.

Correspondingly, the same sample split and analysis was applied to the data below the threshold, as represented in the graph below.

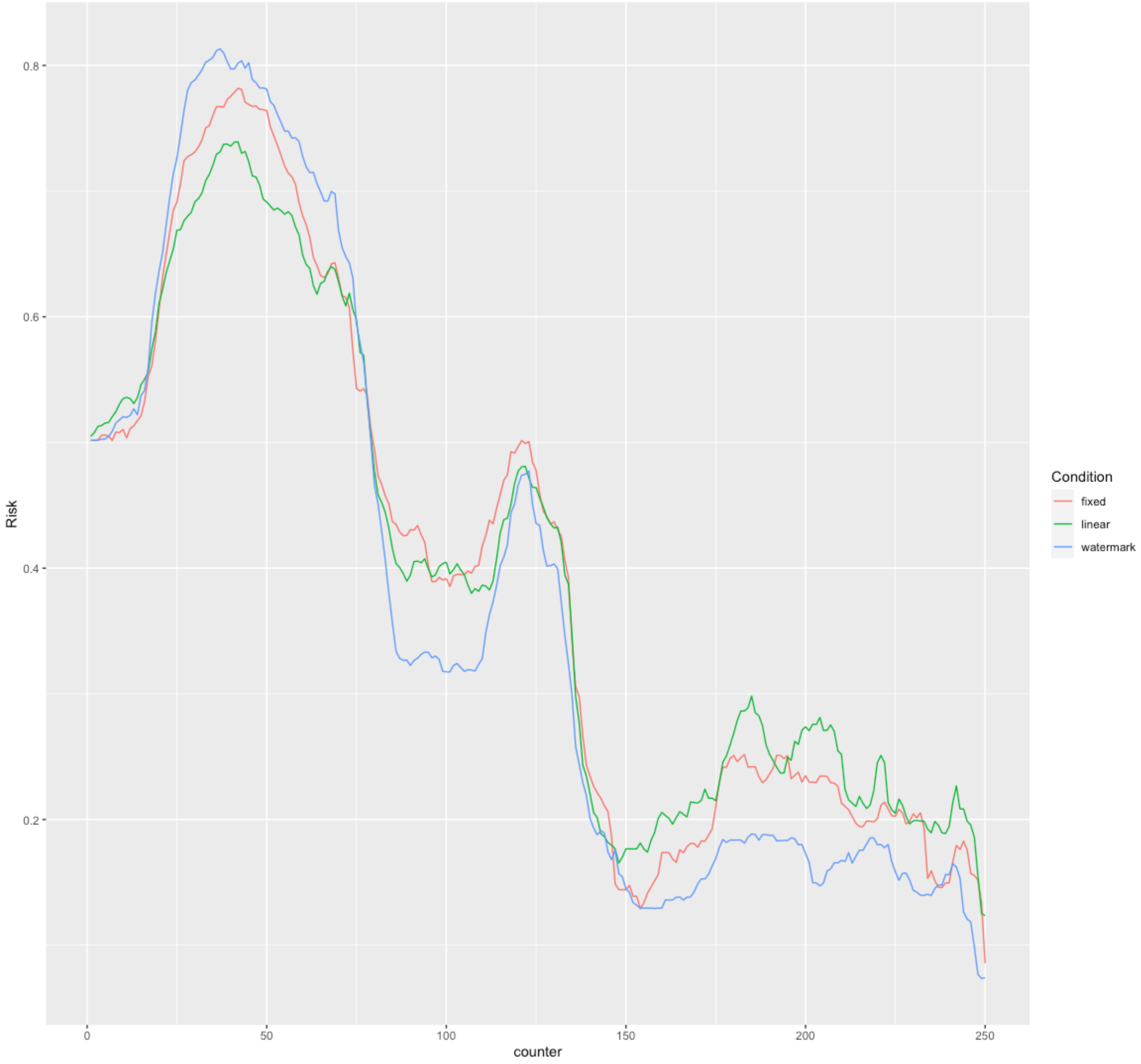

Figure 11: Risk over Days Below the Threshold.

The graph clearly shows that the differences increase; however, it is not evident that as the days go on, the differences grow. This could be because being below the threshold longer might mean either moving further away from or getting closer to the threshold. Watermark seems to have the biggest variance below the threshold, while both the fixed and linear schemes show similar behavior. The next step is to aggregate and compare the conditions using previously applied tests.

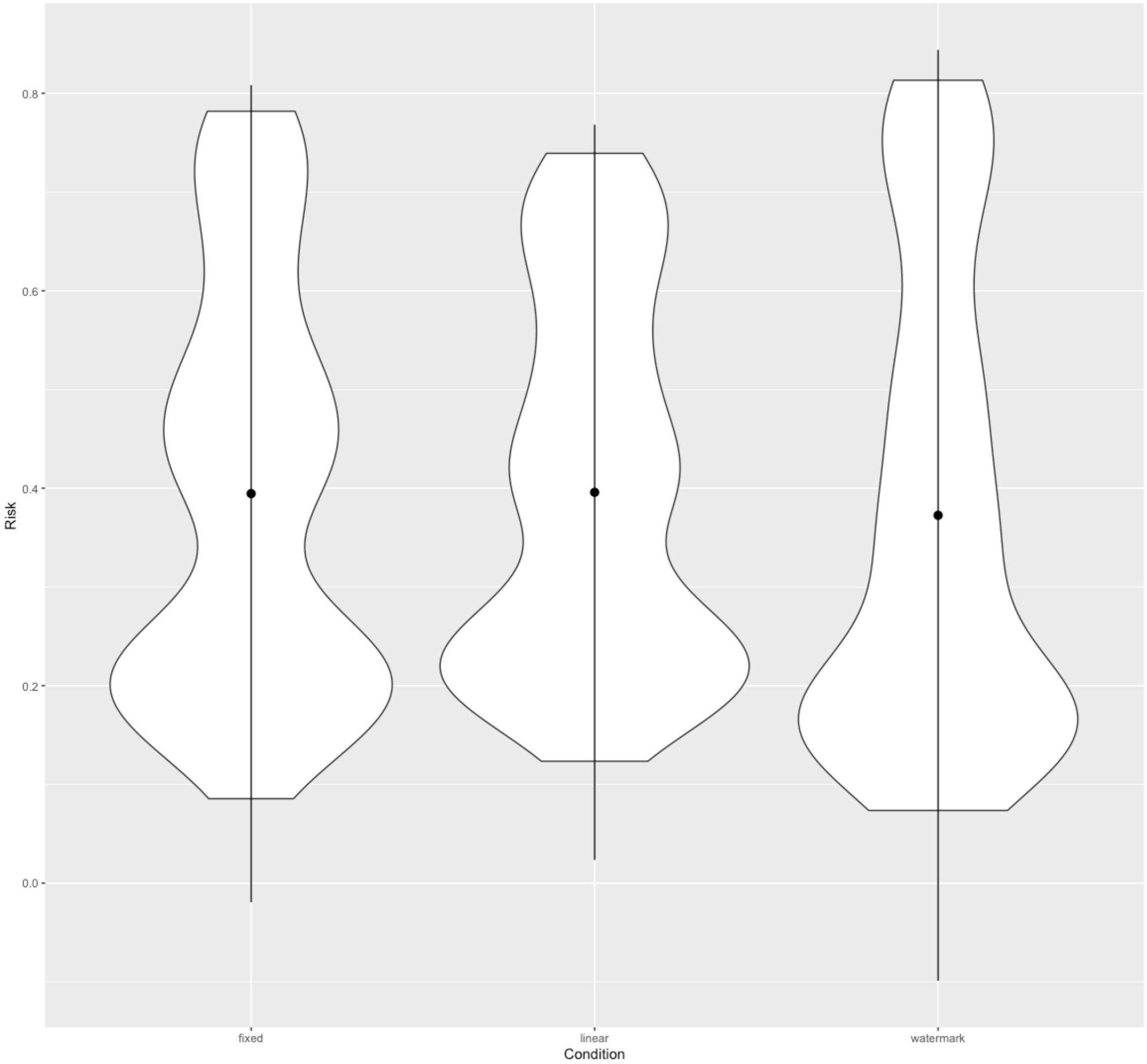

Figure 12: Risk by Condition Below the Threshold (Black dot shows the mean of Risk).

In terms of hypothesis testing, the variance between the three groups differs at a 1% significance level (p-value = 0.002891). The distribution is also statistically significant at 1% (p-value = 0.003279). The difference in means is statistically significant at 5% for the differences between the watermark and fixed groups (p-value = 0.0229), significant at 1% between the watermark and linear groups (p-value =

0.0031), and not significant between the fixed and linear groups (p-value = 0.4767). The three groups behave differently below the threshold, suggesting that trading strategies depend on the compensation scheme. However, the difference is not as strong as above the threshold, making them more susceptible to individual risk preferences.

3.0.1 Additional Tests: Gender-risk differences

According to previous literature, gender differences exist in risk-taking behavior. Men are more likely to engage in risky behaviors (Byrnes et al., 1999). Based on the previous hypotheses, both risk and self-reported risk are not normally distributed. Therefore, the same three non-parametric tests were applied as in previous hypotheses, along with visual data representation. The sample of 201 participants was split into 110 males, 90 females, and 1 non-binary person. The non-binary person was excluded from the analysis due to the sample size being too small to draw meaningful conclusions.

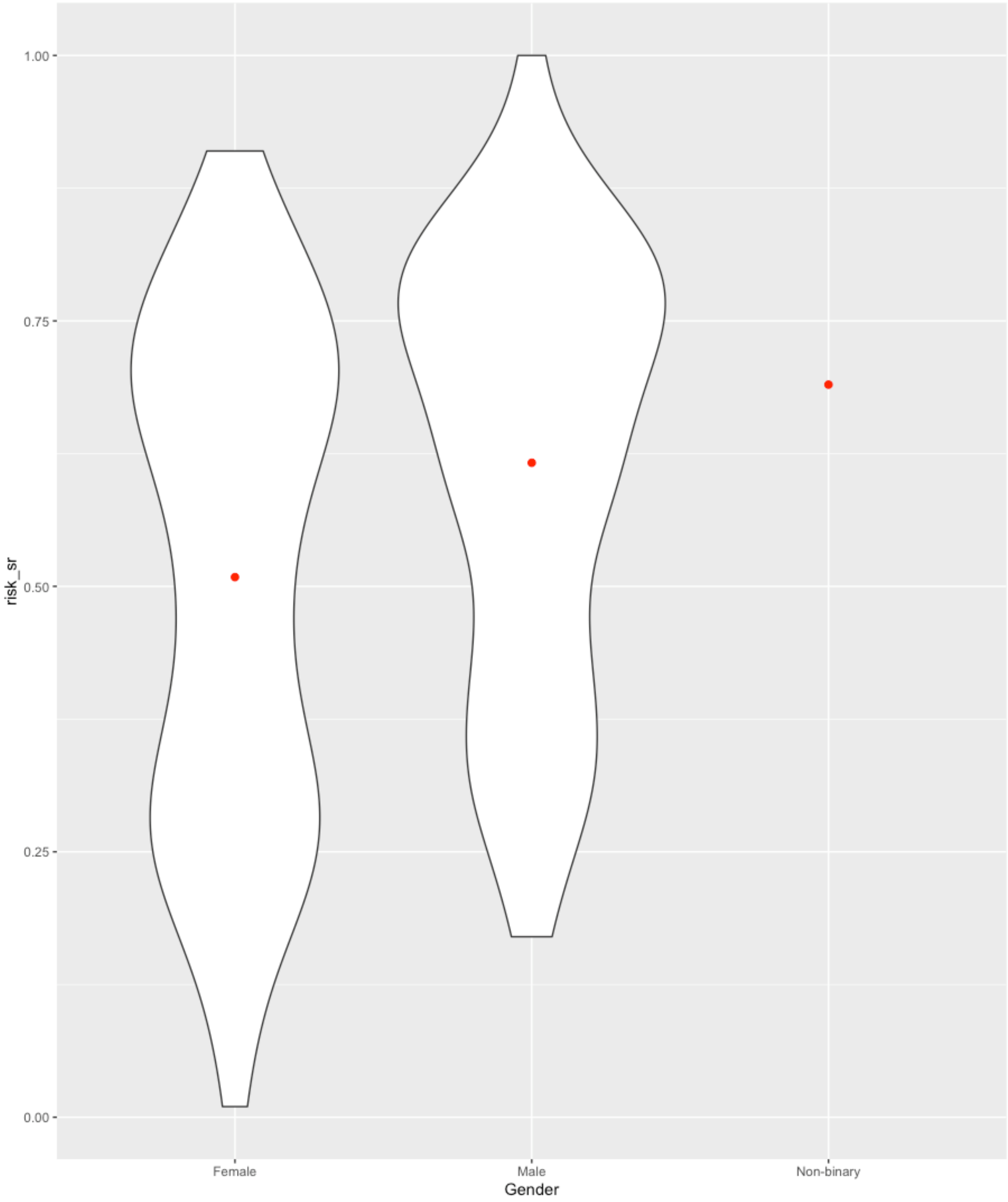

Figure 13: Gender Differences for Self-reported Risk (Red dot represents the mean).

Self-reported risk varies both visually and is confirmed by hypothesis testing. Males and females have statistically significant differences in self-reported risk means at 1% (p = 0.0028). They also have different distributions (p-value = 0.003936) at 1% and variances (p-value = 0.09318) at 10%. The next step is to measure actual differences in in-game risks.

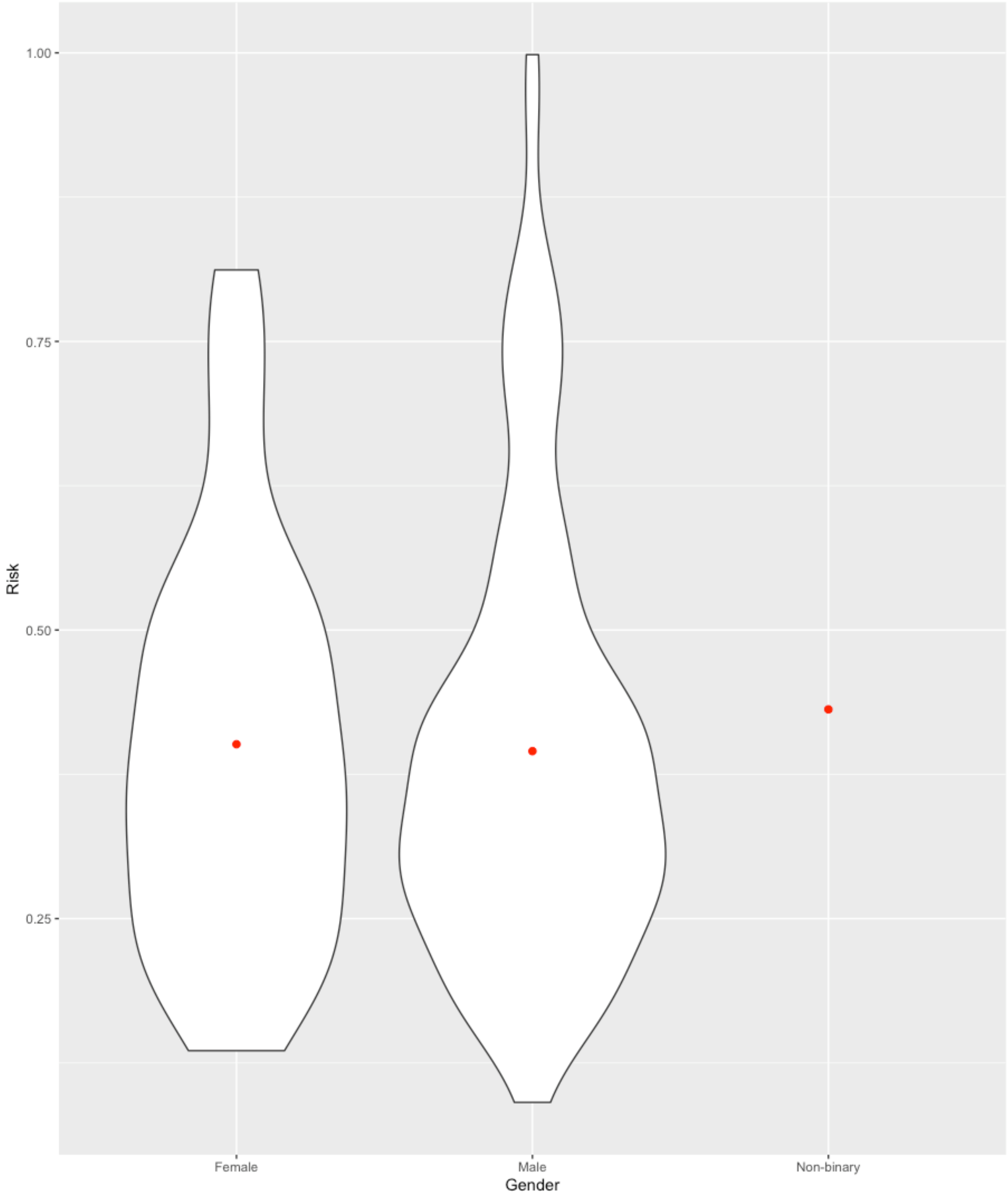

Figure 14: Gender Differences for In-game Risk (Red dot represents the mean).

The in-game risk does not vary based on visual interpretation, and hypothesis testing rejects gender differences. Males and females do not have statistically significant differences in in-game risk means (p-value = 0.003936). They also have similar distributions (p-value = 0.7587) and variances (p-value = 0.409). All the above hypotheses are rejected at the 10% significance level. Knowing that men have higher self-reported risk suggests that the experimental setup had a greater effect on them, reducing their risk profile more than for women. This indicates that gender differences have been mitigated by the experimental treatments.

3.0.2 Return on Investment

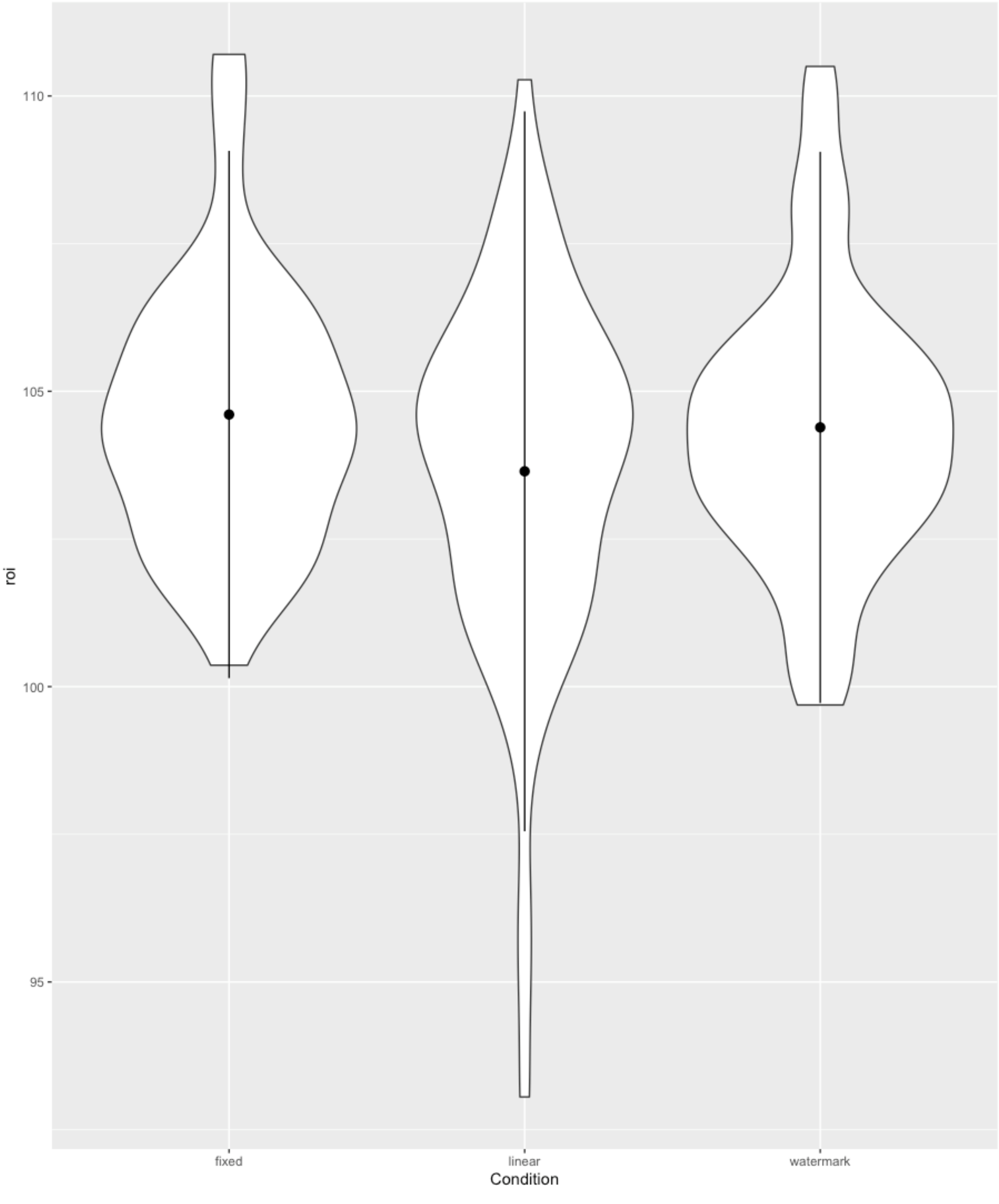

Figure 15: Return on Investment by Condition (The dot represents the median). N=201

Based on the graph, it could be concluded that the linear condition should have a lower ROI compared to the other conditions. However, all statistical tests for variance (p-value = 0.227), distribution (p-value = 0.3051), and mean differences (p-value = 0.39, p-value = 0.66, p-value = 0.46) are not statistically significant at 5%. This means the hypothesis that experimental conditions impact ROI can be rejected. Even though there is no statistical significance, it is interesting that participants in the linear condition show a much broader range of results. This suggests that the manipulation has worked. Participants in the conditions with thresholds focused much more on optimizing their returns around the threshold. This can be seen in both the fixed and watermark violin plots, where most of their distributions are clustered around an ROI of 104 (a 4% increase).

4 Conclusion

The results of this study suggest that the trading behavior of individuals is impacted by compensation schemes. The trading behavior was analyzed in three different dimensions. The first dimension relates to the size, transaction value, and frequency of trades. When comparing the three experimental groups, it was found that the linear compensation scheme resulted in significantly different behavior in terms of the average size of transactions and the average transaction share value, compared to the fixed and watermark schemes. However, there was no statistically significant difference in the number of trades between the groups. The second dimension relates to risk-taking. On one hand, the experiment showed that self-reported risk and elicited risk are not correlated. This has significant implications for the insurance and banking sectors when trying to assess potential risk profiles purely based on survey data. On the other hand, looking at aggregated risk across the experimental conditions, the statistics were not significant, implying no difference between the groups. However, a visual analysis of the risk behavior concerning the price path implied differences at certain time points. The last dimension of analysis is related to threshold behavior. There were significant differences in the experimental condition both below and above the threshold. This implies that the experimental manipulation was successful, and participants attempted to optimize their payoffs according to the compensation scheme.

In terms of expanding the current behavioral finance experimental research relating to compensation schemes, this experiment brings value by examining behavior in a dynamic setting. Differences in risk preferences have been shown in other experiments, but most of them focused on choosing between A and B and analyzing utility curves. The dynamic setting using the Zurich Trading Simulator (ZTS), combined with price randomization, allows for more generalized conclusions. It also provides more insight into relatively under-researched variables, such as the frequency of trades, average trade size, trade value, as well as dynamic threshold behavior.

References

Ackermann, C., McEnally, R., & Ravenscraft, D. (1999). The performance of hedge funds: Risk, return, and incentives. *The Journal of Finance*, *54*, 833–874.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

Andraszewicz, S., Friedman, J., Kaszás, D., & Hölscher, C. (2023). Zurich trading simulator (zts) — a dynamic trading experimental tool for otree. *Journal of Behavioral and Experimental Finance*, *37*, 1–9.

Andraszewicz, S., Kaszás, D., Zeisberger, S., & Hölscher, C. (2022). The influence of upward social comparison on retail trading behavior [OSF Preprint].

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, *12*(1), 19–64.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367.

Carpenter, J. N. (2000). Does option compensation increase managerial risk appetite? *The Journal of Finance*, *55*(5), 2311–2331.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk-taking. *Journal of Economic Behavior and Organization*, *83*, 50–58.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, *87*, 43–51.

Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*(3), 219–243.

de Figueiredo Jr, R. J., Rawley, E., & Shelef, O. (2019). Bad bets: Nonlinear incentives, risk, and performance. *Strategic Management Journal*.

Dousolier, M. (2021). *The impact of compensation scheme on dynamic trading behavior* [University of Basel].

Flannery, T., & Roberts, S. (2021). Agent motivation and principal anticipation: Non-monotonicity, intentions, and other factors. *International Journal of the Economics of Business*, *28*(3), 335–361.

Fligner, M. A., & Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, *71*(353), 210–213.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381.

Frieder, L. (2004). Earnings announcements, order flow, and returns [November 30, 2004]. *Order Flow, and Returns*.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*(5), 379–385.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Jacobs, H., & Weber, M. (2012). The trading volume impact of local bias: Evidence from a natural experiment. *Review of Finance*, *16*(4), 867–901.

Jensen, M. C. (1990). The modern industrial revolution, exit, and the failure of internal control systems. *The Journal of Finance*, *48*(3), 831–880.

Kouwenberg, R., & Ziemba, W. T. (2007). Incentives and risk taking in hedge funds. *Journal of Banking & Finance*, *31*(11), 3291–3310.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lazear, E. P. (2000). Performance pay and productivity. *The American Economic Review*, *90*(5), 1346–1361.

Malkiel, B. G. (1999). *A random walk down wall street: Including a life-cycle guide to personal investing*. WW Norton & Company.

Mayfield, C., Perdue, G., & Wooten, K. (2008). Investment management and personality type. *Financial Services Review*, *17*(3), 219–236.

Panageas, S., & Westerfield, M. M. (2009). High-water marks: High risk appetites? convex compensation, long horizons, and portfolio choice. *The Journal of Finance*, *64*(1), 1–36.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292.

Ritter, J. R. (2003). Behavioral finance. In *The new palgrave dictionary of economics and the law* (pp. 182–184, Vol. 1). Macmillan.

Shapiro, S. S. (1964). Shapiro-wilk test for normality. *Biometrika*, *52*(3-4), 591–611.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review*, *93*(2), 175–179.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Van Horne, J. C., & Parker, G. G. (1967). The random-walk theory: An empirical test. *Financial Analysts Journal*, *23*(6), 87–92.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.