ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Behavioral Economics

Reference points in competition: Effects from an empirical and experimental study [*]

Student: Vytaute Rimkute
Student ID: 588144
First supervisor: Professor J. T. R. Stoop
Second assessor: dr. M. F. M. Hainguerlot
Date final version: April 28, 2022

Abstract

This paper studies the effect of the reference point on performance. A short empirical analysis was conducted using fixed effects regression to measure the effect of the green line on ski jumping performance. However, the findings lacked robustness; therefore, the study employed an extensive pre-registered laboratory experiment in the rank-order competition. The study used slider tasks (Gill and Prowse, 2012) adapted for mobile phones. Subjects were randomly allocated to treatment groups with different types of imposed reference points: i) static (goal-as-a-reference) – delivered before entering the competition, and ii) dynamic – revealed once respondents advance closer to the reference point. No consistent and significant results could be observed for each treatment. Explanatory analyses revealed that recalling the reference point leads to disappointment aversions.

Keywords: reference point, effort tasks, online experiments, disappointment aversions.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

[*] I would like to acknowledge Professor Jura Liaukonyte for inspiring my interest in the empirical analyses. Dr. Agne Kajeckaite was instrumental in supporting me to follow a experimental economics path. For this, I am extremely grateful.

Contents

1 Introduction **3**

2 Literature **5**

2.1 Prospect Theory and Reference Point 5

2.2 Goals and Reference Points 7

2.3 Competition and Reference Point 10

3 Empirical Findings **13**

3.1 Data and Methodology 14

3.2 Empirical Results 15

3.3 Empirical Conclusion 17

4 Experimental Evidence from Laboratory **17**

4.1 Mobile Slider Task 18

4.2 Experimental Procedure and Hypothesis 19

5 Analysis **22**

5.1 Devices 23

5.2 Descriptive Statistics 25

5.3 Tests on Reference points 27

5.4 Tests on Noticing Reference Point 30

5.5 Regression Analysis 30

6 Discussion and Limitations **35**

7 Conclusion **38**

Appendices **45**

1 Introduction

According to empirical and experimental studies, subjects can be influenced by the reference point. Most of the studies stemmed from psychology and were adapted to the prospect theory by behavioural economists Kahneman and Tversky in 1979. The reference points transform to goals (Heath et al., 1999). Hence, goal-as-a-reference point can be observed everywhere in our daily lives, from personal goals such as losing weight to adjusting the firm performance based on the competitors.

The literature on reference points in economics is extensive. It dates back more than 40 years, when Kahneman and Tversky (1979) proposed the prospect theory, demonstrating that individuals perceive losses and gains differently depending on their position relative to the reference point. Individuals behind the reference point are risk-seeking, whereas those ahead of the reference point are risk-averse. Following this, the reference point can be perceived as a goal (Heath et al., 1999), for example, as status quo (Kahneman and Tversky, 1979), expectations (Ko˝szegi and Rabin, 2007), social comparison (Bolton and Ockenfels, 2000), and others. The goal-as-a-reference-point may be affected by the characteristics of the reference point and the individual's position toward the goal, experiencing the goal-gradient effect. According to Wallace and Etkin (2017), for specific goals (e.g., losing 6 pounds), individuals base their reference point at 6 pounds; whereas, in the "do-your-best" strategy, individuals perceive their reference point at the beginning, and therefore, measure their progress based on how far they've come from the beginning. Finally, one could argue that high stakes, experience, and competitive environments would invalid the prospect theory (List, 2003; List, 2004; Levitt and List, 2008), and therefore, agents would not adjust their behaviour based on the reference point. Gill and Prowse (2012) investigated the competitive environment and proposed that being aware of competitors' accomplishments may lead to disappointment aversions.

My study aims to combine the goal-as-a-reference point and competitiveness in a rank-order tournament by analysing empirical and experimental data. The study first analyses the behaviour of ski jumpers in World Cup tournaments from 2010 until the 2021 season. In 2014, the ski jumpers were given a "to-beat line," which is displayed for each jumper based on the previous best performer in an event. The line reflects the reference point obliquely. I discovered robust negative effects on judge points using fixed effects regression. This presumably implies that athletes began to engage in risk-seeking behaviour once the line was introduced. However, the results were inconclusive because no effect on delta distance was found. In addition, the study lacked strong counterfactual evidence. Therefore, inspired by ski jumping, I conducted a laboratory experiment on the Prolific platform.

The pre-registered experiment employed a between-study design and explored rank-order tournament competition. Agents were randomly assigned to one of three different treatments: i) no reference point, ii) static reference point - individuals were informed about the results from the pilot study upfront, and iii) dynamic reference point - agents received information on the reference point once they were close to the reference position. Their effort towards the reference point was measured based on a mobile slider task, similar to Gill and Prowse (2012). To the best of my knowledge, this is the first study to apply the slider task to the mobile environment. Two types of analyses were carried out: i) non-parametric tests and ii) an OLS regression estimates.

There was no evidence of an effect between the different treatments and the baseline results based on pre-registration. Heterogeneous device effects hampered the estimates, so the analyses were performed separately. The non-parametric tests revealed no differences between the three groups. The difference between static and baseline reference points was slightly higher but not statistically significant. Furthermore, the OLS regression measuring the effect of the treatment on performance was estimated separately by devices. Evidence suggested that individuals assigned to the static reference point improved their performance compared to the baseline condition, but the results were not significant or robust. Dynamic treatments had no effect on the results.

The respondents were also asked if they remembered the reference point. As a result, some explanatory analyses were performed while controlling for it. Individuals who noticed the reference point and competed in the static treatment showed disappointment aversion compared to those who competed in the dynamic reference point condition. The decrease in performance based on correctly moved sliders sought 5.6 points at the 5% significance level, ceteris paribus. This was also economically significant result, as it leads to a 31% increase in performance.

The results of the study contribute to the existing literature. First, although various reference points have been extensively researched, the competitive environment surrounding goal-gradient effects has received little attention. Second, this is the first empirical analysis investigating the green line effect on ski jumping performance. Third, because behavioural laboratories' usage has been limited due to COVID-19, mobile effort tasks contribute to innovative approaches to measure agents' efforts while ensuring homogeneity.

In the remainder of the paper, section 2 presents the contributing literature. Section 3 provides empirical analyses from ski jumping and explains the link between the empirical data and the laboratory experiment. Section 4 describes the experimental design and procedure. Section 5 evaluates the results. Section 6 discusses the limitations; the conclusion is presented in section 7.

2 Literature

In the literature review, I aim to provide baseline knowledge of prospect theory to explain the origins of reference points. Furthermore, I explore examples from laboratory and field experiments to discuss the goal-as-a-reference point. Lastly, I discuss the goal-setting upon the reference points in a competitive environment.

2.1 Prospect Theory and Reference Point

Prospect theory, introduced by Kahneman and Tversky in 1979, is a violation of expected utility. In 1992, Kahneman and Tversky expanded the theory to include a cumulative prospect theory. The initial paper has been cited over 70,000 times (Econometrica, n.d.), so the theory has evolved and is now applied in various fields. In this section, I will try to summarise the most important aspects of prospect theory and the reference point related to my study.

The reference point stems from prospect theory (Kahneman and Tversky, 1979). Individuals in the laboratory consistently violate expected utility, according to Kahneman and Tversky (1979). This is a standard theory used by economists when making decisions under risk. The prospect theory comprises three building blocks: i) loss aversion, ii) diminishing sensitivity, and iii) reference point. In 1992, Kahneman and Tversky introduced the fourth block: iv) probability weighting. First, individuals overestimate losses in comparison to the gains, resulting in loss aversions. Second, agents experience diminishing sensitivity, thus each additional gain further from zero value leads to a lower marginal utility. Therefore, agents seek a specific reference point to maximise utility while avoiding loss. Third, the reference point distinguishes between loss aversion and diminishing sensitivity values. Lastly, probability weighting is related to tail outcomes in any distribution, causing individuals to overweight unlikely outcomes (Barberis, 2013).

Prospect theory has a graphical implication, with the most well-known feature being a discontinuity or kink at a reference point, identifying loss aversion. This is encountered by either the discontinuous first derivative of utility at 0 or as a "kink" in the utility before the value 0 or the reference points (Kahneman and Tversky, 1992). Therefore, when analysing individual behaviour at the reference point, studies often provide a value distribution, and any sudden change around the reference point is interpreted as the existence of loss aversion. Allen et al. (2017) conducted an empirical study on marathon runners who set their goals with rounded numbers (e.g., 4 hours, 4:30, etc.). According to the study, runners' finishing time distribution before the rounded numbers depict a "jump," resulting in runners being faster at the reference point than those finishing after. Therefore suggesting that individuals experience loss aversions and rounded numbers as a goal-as-a-reference point for them.

One of the main challenges of applying prospect theory is the lack of precision. Barberis (2013), in his overview, "Thirty Years of Prospect Theory," discusses that it is still unclear how we define "losses" and "gains" and determine the reference point. Ko˝szegi and Rabin (2007) rises the idea that the reference point is based on the initial value, the so-called status quo, and that the gains and losses can be computed based on expectations. An example of the endowment effect was provided, where randomly assigned mug owners in an experiment (Kahneman, Knetsch, and Thaler, 1990) were willing to sell the mug at a higher price than nonowners. Therefore, mug owners' willingness to sell was based on an expected reference point, while nonowners' willingness to sell was based on the status quo. Clearly, the academic literature over the last 40 years has improved the theory and implications of prospect theory. The majority of studies rely on positive economics and attempt to explain the observed behaviour. However, some studies focus on prescriptive application, thus nudging individuals to the preferred behaviour (Barberis, 2013).

To summarise, after 40 years of prospect theory, economists and psychologists promote research and application in various fields. Prospect theory, which originated as a decision under risk, has found applications in finance, insurance, consumption and saving decisions, industrial organisation, labour supply, and other fields (Barberis, 2013). Reference points can be classified into different types based on various criteria. However, there is a significant distinction between internal and external reference points (Bell and Bucklin, 1999). In a literature review, Wang et al. (2020) assign expectations, goals and aspirations, minimum requirements, social comparison, and status quo to internal reference points. On the other hand, external reference points are externally stimulated and based on previously available information, such as goals and rewards (Wang et al. 2020).

2.2 Goals and Reference Points

Goals can be perceived as a reference point (Heath et al., 1999). Studies combined goals and reference point literature. Furthermore, they explored various sources influencing goal achievement, such as the agent's position in the progress curve (Cheema and Bagchi, 2011; Koo and Fishbach, 2012), visual ability (Cheema and Bagchi, 2011), impact level (Nunes and Dreze, 2006), and others.

Heath et al. (1999) merge prospect theory with the goal-setting theory and thus conclude that reference point can be seen as a goal-setting mechanism. Agents do not perceive outcomes as neutral; they categorise them as failures or successes. Therefore, goals are rather a deviation between these two outcomes. Additionally, Heath et al. (1999) demonstrate that both effort and satisfaction outcomes can explain the prospect theory values. Heath et al. (1999) provided the following example about Charles and David on effort surveys: Charles sets a goal of 30 sit-ups, whereas David seeks to achieve 40 sit-ups. Charles and David are both on their 34th sit-up, and thus students [N = 74] are now asked who would exert more effort for one additional sit-up even though both are exhausted. Study subjects suggested that David (82%) would invest more effort than Charles (18%), which means that an additional sit-up for David yields higher utility in terms of satisfaction than for Charles. Agents believed that Charles achieved the goal-as-a-reference point; however, David was still behind, so he was more likely to exert more effort to avoid personal failure. In other words, this behaviour is perceived as experiencing loss aversion. Consequently, Heath et al. (1999) state that the agents value the goal-related outcomes (i.e., levels of goal progress) relative to the reference point, which further leads to a discussion on the value function.

The robust findings in the literature of goal achievement is the strong relationship between effort, motivation, and goals (Hull, 1932; Fishbach et al., 2010; Koo and Fishbach, 2014; Harkin et al., 2016). The "goal looms larger" or "goal-gradient" effects apply to individuals who accumulate and exert more effort as they get closer to the end. The desired goal by Heath et al. (1999) serve as a reference point, and goal pursuits are posited as a value function. The function is driven by effort; therefore, each increase in value leads to a higher marginal utility when approaching the goal. In other words, each increase in effort has a greater impact on the overall goal, resulting in greater achievement (Wallace and Etkin, 2017).

Furthermore, goal specificity may result in different outcomes. Cheema and Bagchi (2011) conducted a laboratory experiment to investigate the effect of visualisation on effort exertion. Undergraduate students [N = 79] were asked to exert effort using a hand dynamometer at a consistent level for 130 seconds (130 seconds corresponds to 4.33 rotations of the watch hand on a 30-second stopwatch). The study included two treatments: easy-to-visualise and hard-to-visualise groups. In the easy-to-visualise group, students saw a horizontal bar on a screen that filled in as time elapsed. Alternatively, the hard-to-visualise group was presented with a 30-second stopwatch. Consequently, in the hard-to-visualise condition, the clock was updated each second, whereas, in the easy-to-visualise condition, the progress of incrementally filling the bar was seen. One would argue that the difference between the treatment conditions is relatively small, but the results presented in four stages provide evidence of a significant effect in effort exertion for the last 30 seconds. The visualisation did not affect the first three stages. Nevertheless, individuals were mostly affected by the goal visualisation once they approached the end.

A supporting experiment was conducted by Koo and Fishbach (2012). They explored the difference between high and low goal attainment progress with accumulated and remaining progress fulfilment. In the laboratory experiment, participants from the University of Chicago were provided with partly completed frequent coffee cards. The card would be used to collect ten stamps to receive a free hot beverage as a reward. Researchers employed 2 x (focus: remaining vs. accumulated progress) and 2 x (progress: low and high) between-subject design. For example, the manipulated groups would have three or seven accumulated or remaining stamps on their card. The study measured outcomes by the willingness to enrol in the programme. The results of the high progress treatment were significantly more motivating for students to achieve the reward than the results of the low progress treatment. In the accumulating treatment, there was no difference between low and high progress. However, participants indicated that the remaining fulfilment of the goal is more motivated in the later stages than in the earlier ones. Lastly, Cryder, Loewenstein, and Seltman (2013) used an empirical study of crowdfunding campaigns to support the goal-gradient effect. Three studies showed that the donors' charitable actions increase as the campaign gets closer to its goal, suggesting that donors perceive their impact as higher when approaching the end.

Cheema and Bagchi (2011) conducted another study in which they measured the effect of progress on goal attainment in a 2x2 between-subject design experiment. Students [N = 183] were divided into two groups and given either a difficult or easy goal for saving \$750. They were told they had \$225 (goal far) or \$525 (goal near) and were thus randomly assigned to either the visualised bar filling treatment (goal easy) or the text treatment (goal difficult). After the participants completed a survey measuring their savings commitment, the results indicated that those who were closer to their goals were more committed. Furthermore, the visualisation of the goal in the easy condition increased the likelihood of goal achievement. On the other hand, visualisation had no effect on the outcomes in the difficult goal treatment.

Wallace and Etkin (2017) investigate the distinction between specified and non-specified goals. Goals like "lose 6 pounds" are specific, whereas "do your best" is not. The absence of a specific end goal introduces ambiguity in performance evaluation and may even result in poorer results (Wright and Kacmar, 1994; Clark et al., 2020). Nevertheless, it begs the question of how people perceive non-specific (vs. specific) goals and the position of the reference point. Wallace and Etkin (2017) try to answer the question by hypothesising that the lack of a specific goal induces the reference point at the outset. Individuals with non-specific goals value progress based on their starting point. Alternatively, as previously discussed, an end reference point induces the specific goals. Wallace and Etkin (2017) propose a theory that non-specific goals are driven by a diminishing sensitivity function and thus should be steeper in terms of motivation values at the beginning of the goal progress. That is, as individuals move away from the initial reference point, their progress appears less significant, and as a result, they experience diminishing sensitivity. On the other hand, specific goals are motivated by loss aversion and thus have a steeper slope at the end. As discussed in a previous example, agents experience goal-gradient effects and, as a result, exert more effort closer to the end-stage than at the beginning. At the behavioural laboratory, the researchers conducted an experiment with 155 students. Participants were assigned to 2 x (goal: specific vs. non-specific) and 3 x (progress: low, medium, and high) between-subject design treatments at random. Students were instructed to find an error in the text passages, which each contained an error. Furthermore, finding each error in a passage would result in a streak. Specific goal treatments were instructed to find at least ten errors, whereas non-specific treatments were instructed to find "as many as possible." Failure to find an error would bring the streak to an end, and thus the game would be lost. The study was paused after finding two errors, five errors, or both errors, simultaneously corresponding to low, medium, and high treatment conditions, and individuals were given new instructions stating that the next passage would be more difficult and that if they could not find the error, they would be asked to leave the study. However, because there were no errors in the passage, people would eventually leave. The time spent searching for the error serves as a proxy for motivation. According to the hypothesis, researchers discovered that non-specific goals provided significantly less motivation than specific goals. Individuals with non-specific goals were also more likely to invest progressively low effort from a low to high progress condition.

To summarise, the specificity and framing of the goal-as-a-reference point may produce different outcomes. Individuals exert a goal-gradient effect closer to the endpoint, whereas they may be behaving according to a diminishing sensitivity curve in the early stage. Nonetheless, everybody is surrounded with a context, raising new questions about how a competitive environment affects the reference point outcomes.

2.3 Competition and Reference Point

Fiegenbaum, Hart, and Schendel (1996) developed the strategic reference point theory by combining prospect theory and organisations. They argue that when a company is above the reference point, it is risk-averse, whereas it is risk-seeking when it is lagging behind the reference point. According to the authors, the strategic reference point matrix includes internal, external, and time dimensions. External success is most commonly defined as outperforming the competition and thus leading the market. Thus, actions in a competitive market are defined in relation to the competitors (Shoham and Fiegenbaum, 1999). Therefore, competitor outcomes serve as a reference point and are endogenous to agent performance. Furthermore, competition involved various behaviours, which led to the development of the contest theory (Konrad, 2009).

Players in the contest can exert scarce resources such as effort, motivation, and money to increase their probability of winning the competition (Konrad, 2009). In the contest, each agent is ranked individually; thus, the best performer receives the prize. The prize in the contest is directly influenced by the outcomes of other players, because agents must exert more effort than the closest competitor in order to be a leader (Dechenaux, Kovenock, and Sheremeta, 2015). In real life, various forms of contests can be observed, such as all-in-auction or Tullock contests, but for our purposes, we are primarily interested in rank-order tournaments. Individuals in rank-order tournaments are ranked from best to worst in terms of the task on which they are competing (Konrad, 2009). The best performer receives the top prize, the second-best performer receives the second-highest prize, and so on. Rank-order tournaments can be found in sports competitions (Fershtman and Gneezy, 2011), salesperson performance (Casas-Arce and Martinez-Jerez, 2009), and promotion execution (Bognanno, 2001).

Contests suffer from heterogeneous effects across agents. The number of players (Lim et al., 2014), risk attitude (Cason, Masters, and Sheremeta, 2020), gender (Niederle and Vesterlund, 2011), prize type (Fairburn and Malcomson, 2001), and other factors might lead to different outcomes in a tournament. The contest theory assumes risk-neutrality (Konrad, 2009). However, it is well researched that different risk attitudes lead to different outcomes. Cason, Masters, and Sheremeta (2020) conducted a laboratory experiment to analyse the effort exerted in three winner-take-all contests with different types of payoff distribution. The findings suggested that the overall exerted effort differs between the contests; however, subjects' risk attitudes towards the goal remain consistent across all contests. Risk-averse individuals tended to exert less effort in tournaments, whereas risk-seeking individuals were more competitive. Theoretically, agents who believe they are far ahead in a large number of participants' competitions slack off because they believe losing is impossible. Alternatively, performers who are significantly behind give up because catching up seems impossible (Casas-Arce and Martinez-Jerez, 2009). However, the Casas-Arce and Martinez-Jerez (2009) study provided no empirical support for the latter scenario, possibly because of the research's attrition bias. However, the quitting behaviour is in line with giving up in the competition. Fershtman and Gneezy (2011) conducted a field experiment in four schools in which 10th grade high-school students ran a 60-metre race during a physical class. The first run was done individually, whereas the second run was done in pairs, with partners chosen at random or based on their ability. Pairs could also run together or separately. Lastly, low, middle, and high-level incentives were introduced for the competitions. The results suggested that the quitting behaviour correlated with the incentives. The quitting rate was only significant for the high level of incentives. Fershtman and Gneezy (2011) propose a theoretical model in which high-level incentives imply high-level effort, resulting in a higher cost of finishing the competition. They posit that because the higher cost of finishing is greater than the social cost of quitting, agents prefer to withdraw from the competition. Disappointment aversion is also defined as putting in less effort, motivation, or other performance-based values (Gill and Prowse, 2012).

The discouragement effect appears under the shared information of competitors' outcomes (Gill and Prowse, 2012). According to research, subjects perform differently depending on whether they trail or lead their competitors and experience disappointment aversion (Gill and Prowse, 2012; Ludwig and Lünser, 2012; Eisenkopf and Teyssier, 2013), suggesting that the competitor's performance serves as a reference point. Gill and Prowse (2012) designed a two-stage real effort laboratory experiment and paired subjects to perform a "slider-task." The game consisted of sliders that must be moved to a specific position. The agent who moved the most sliders in a limited time in both stages wins the prize. Individuals were randomly assigned to "First movers" and "Second movers". Once the "First mover" completed the task, the "Second Mover" observed the results and started the game too. Assuming that subjects were focused on the prize, the "First Mover" effort level should have no effect on the "Second Mover." Nonetheless, the results showed that when the "First Movers" exerted a high level of effort, the "Second Mover" shied away and became less motivated. Alternatively, if the "Second Mover" encountered low effort from the "First Mover," she exerted more effort to reach the baseline. This demonstrates that subjects in competitive environments are not only motivated by monetary gains, but also adjust their goals in real-time based on the competitors' efforts, which serves as a reference point.

Field experiments yielded similar results. Delfgaauw et al. (2014) investigated whether providing performance-based information to stores would help them meet their sales targets and earn bonuses. Researchers collaborated with an entertainment retail store chain in the Netherlands, which implemented a new bonus system to boost sales performance. A total of 189 stores were randomly assigned to either a treatment [N = 93] or control group. The treatment group was given weekly information on the performance of the control group. Employees in the treatment condition only received bonuses if their stores outperformed comparable stores in the control group. No bonus scheme was provided to the control group. The findings indicated that the dynamic incentive scheme had no effect on sales performance on average. Perceived information positively affected sales performance in stores that were closer to the goal. There was no response to the incentives found for the stores that were far behind. As in previous studies, it was possible that employees gave up and, as a result, experienced disappointment aversion.

In summary, the reference point stems from prospect theory. Various laboratory and field experiments provide evidence that individuals made decisions based on the reference point. Due to goal-gradient effects and risk-seeking behaviour, they exert more effort by being closer to the reference point. Once the reference point is behind them, agents are risk-averse. The goal-as-a-reference point can range from something salient, such as the green line in ski jumping, to anticipated competitor performance in competitions. Further in my study, I explore both and attempt to link visual reference point positions in the competition, which challenges study participants to anticipate their competitors in the rank-order tournament.

3 Empirical Findings

Empirical findings from ski jumping inspire the researcher to investigate the reference point. Ski jumping has grown in popularity in Central Europe over the last 20 years (FIS-ski.com, 2018). Changes were made regularly to increase TV audience engagement. One of these improvements was the introduction of the green line, also known as the “to-beat line” (FIS-ski.com, 2014). The green line was projected on the ski jumping hills based on the last best performance and thus identifying the jumpers that overcome the line would increase their chances of being leading athletes in the competition.[1]:

The performers jump in the order specified from the worst to the best based on qualification results. On average, agents should outperform the line in terms of distance in line with Dechenaux, Kovenock, and Sheremeta (2014) findings. A rational agent should not be influenced by the line and thus jump as far as possible to win the competition, because the green line provides an asymmetric information. However, it is possible that agents deviate from the expected utility and thus perceive the green line as a reference point. Performers would exhibit risk-seeking behaviour when ahead of the line because they might lose, and they may be risk-averse when after the line. Prospect theory (Kahneman and Tversky, 1979) and the goal-gradient effect support such a behaviour (Koo and Fishbach, 2014; Harkin et al., 2016; Wallace et al., 2018). Lastly, the experimental findings from Cheema and Bagchi (2011) suggested that

[1] The line is clearly visible to the performers. An active ski jumper R. Kobayashi in one of his interviews mentions “<..>after the take-off you see the green line and if you made it or not.” (Perelman, 2019)

the visibility of the goal increases the probability of achieving it.

3.1 Data and Methodology

In order to measure the effect of the reference point on performance, the ski jumpers performance was analysed. The data was retrieved online from the International Ski Federation statistics and contains 9270 individual jumps data of the first-round jumps [2]. In 2010, compensation points for wind and gates were introduced. The change had an effect on the performance scores and, presumably, the jumping strategy (Virmavirta and Kivekäs, 2012), so only the data from 2010 was analysed. Apart from the green line presented in 2014, no other significant changes, to my knowledge, have been introduced between 2010 and 2021. Therefore, I can use a fixed effects regression to compare the outcomes before and after the introduction of the "to-beat line." In this case, the main outcome after the green line was introduced is compared to the time before the line. The regression equations is:

$$y_{it} = \beta Green\,line_t + \gamma Wind\,speed_i + \phi_{athlete} + \delta_{order} + \eta_{competition} + \zeta_{hill} + \varepsilon_{it} \tag{1}$$

The outcome was the delta variable measuring how far the individuals i in year t landed from the last best performer[3]. Green line is a dummy variable that indicates 1 if the green line was displayed in the competition from 2014 and 0 if years before. The β is the coefficient of interest, which represents the change in delta after the green line was introduced in the tournaments relative to the period before. The wind speed control was introduced because it is an outdoor sport, and wind plays a significant role in the competition (Virmavirta and Kivekäs, 2012). The equation includes fixed effects for athletes, order, hill size and competition to account for unobserved unit heterogeneity between these characteristics. Adding fixed effects helps us to control for systematic differences in outcome variables across athletes, order, hill size and competition. More precisely, it increases the probability of comparing akin athletes in terms of order, hill and competition before the introduction of the green line and afterwards.

[2] The first round data was used due limitations in converting data from PDF to readable format. Presumably, the second round data would be less random, because athletes jump in order of the first round results. The first round results are based on the qualification results happening few days before.

[3] Even though there was no line before 2014, I could reconstruct the line, by calculating where it would have been based on the last best performer. One could refer to it as a placebo line.

The same equation (1) was performed on judge points. Measuring the effect of the green line on judge points could imply risk-seeking behaviour. Judge points evaluate the jumpers' flying style in terms of flight, outrun and landing (ICR, 2021). I expect that without the salient line, individuals were not aware of their exact landing position and would have landed in order to maximise their judge points. Once the line is introduced, athletes, while landing, can see the line and thus might try to overfly the line, but risk the landing position. This would have a negative effect on judge points. The standard errors were clustered at event level.

3.2 Empirical Results

Table 1 shows the effects of the green line on jumpers' performance in terms of the jump delta to the line position and judge points. Column 1 shows a reduced form regression of the green line effect on performance while controlling for wind speed. The effect was negative, but it was insignificant. Figure 1 shows a graphical extension of the results in Column 1, but with the jump order dummy, and we could see that the green line had no effect on delta because the confidence intervals overlapped. However, the figure suggests that there may be heterogeneity in jump order because the confidence intervals overlap from the 30th jumper in order, but not as extensively as before. The effect could be more pronounced on the first 30 jumpers because they have lower skills, and having a goal-as-a-reference line may be more effective than not having one. Individuals who land above the line fly further than those who land below the line because the delta was positive. Column 2 only shows the results for the first 30 jumpers; the sign had changed in comparison to the overall effect, but the effect was still insignificant. Column 3 presents the results from the equation (1), yet, no significant effects could be found either. One could argue that the green line had no effect on the ski jumpers and that having a reference displayed had no effect on their performance. However, the heterogeneous effects of the jump order and the lack of a good counterfactual may be impeding the results. Following that, delta had been calculated using the distance measure, which lacks precision because it is rounded to 0.5 (ICR, 2021).

Risk-seeking behaviour could be captured by judge points. Figure 2 shows that having a green line reduced the judge points and substantially increased the overlap of the confidential intervals after the 30th jump. The overall effect estimates were displayed in Column 4, providing evidence that having a green line, compared to the period without the line, decreased the judge

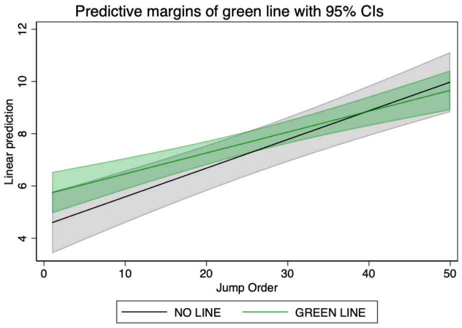


Figure 1: The effect of the green line on the delta.

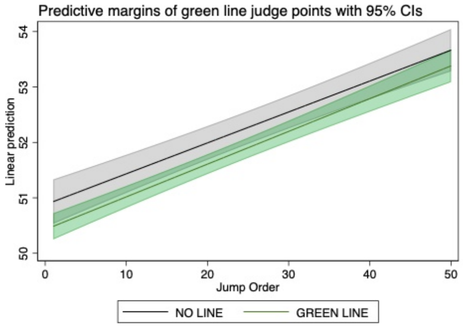


Figure 2: The effect of the green line on judge points.

points at the 5% significance level, ceteris paribus. Column 5 only accounts for the first 30 jumpers, so the effect was more pronounced. Lastly, Column 6 presents the results of equation (1). According to the estimates, having a green line versus not having a line reduces the judge points by a score of 0.46 at a 5% significance level, ceteris paribus. It is worth noting that the effect was not economically significant, as the relative decrease was 0.0089% (Mean = 52.03). Yet, these consistent and negative estimates suggest that the “to-beat line” may reflect the reference point, but the interpretation is limited due to confounding effects and low measure precision.

Table 1: The effect of the green line on performance

	(1)	(2)	(3)	(4)	(5)	(6)
	Delta	Delta	Delta	Judge points	Judge points	Judge points
Greenline	-.227	.096	.763	-.163**	-.29***	-.465**
	(.2)	(.279)	(.643)	(.067)	(.081)	(.232)
Wind speed	.059	.056	-1.39***	.24***	.252***	.77***
	(.111)	(.151)	(.296)	(.039)	(.046)	(.117)
Observations	9270	5480	5434	9459	5669	5631
R-squared	0	0	.226	.005	.007	.216
First 30	NO	YES	YES	NO	YES	YES
FE	NO	NO	YES	NO	NO	YES
Cluster	NO	NO	YES	NO	NO	YES

Note: Cluster Standard Errors on event ID had been added to 3 and 6 columns. Other equations were performed with robust standard errors. Standard errors are in parentheses *** p<.01, ** p<.05, * p<.1

3.3 Empirical Conclusion

One could argue that the identification strategy used to observe the causal effect was insufficient. Including fixed effects may improve estimates of the relationship, but it does not imply causal inference. Ski jumping was constantly improving, and even when athlete fixed effects were taken into account, seasonal unobserved athlete development could not be captured. Ideally, a more efficient counterfactual was required. For example, using the ski jumpers performance from a continental cup, where the green line is not displayed. Summarized, any random allocation of individuals jumping with and without a green line in the similar type of competition and hill size would allow one to capture a causal effect.

Nonetheless, the results suggested that there is a correlation between the green line and performance, but various confounds hampered the results. The green line was accounted for as a reference point based on significant negative judge points. These findings provide information for further research and, as a result, lead to the laboratory experiment.

4 Experimental Evidence from Laboratory

The pre-registered experiment was conducted on a mobile phone using the Prolific platform. The study employs an effort slider task similar to that used by Gill and Prowse (2012). Subjects were asked to compete in a rank-order tournament, moving as many sliders as possible in the limited time they had on their phones. The study uses a between-subject design, with participants randomly assigned to one of three groups: baseline, static, or dynamic reference point condition. The experiment is divided into three stages: instructions, effort task/tournament, and questionnaire, and it was designed with Qualtrics. In the absence of a real laboratory environment, it is critical to ensure homogeneous devices for measuring the exerted effort. Prolific allows researchers to publish surveys on a specific device, so the experiment is only available for participants using a mobile phone. Furthermore, only subjects from the United States are invited, as the majority of the population in the United States owns an iPhone (Statista, 2021), increasing the probability of homogeneity in the study.

The experiment held two sessions with 172 participants. Subjects were paid a £0.38 flat fee, and the TOP3 participants shared a prize of £8.89 (12$). Six participants received a prize of

£1.47 as a result of the tie. The session took around four minutes[4]. This leads to hourly average earnings of £5.7, which was above the lowest payment on Prolific.

4.1 Mobile Slider Task

Gill and Prowse (2012) developed the original slider task, and it is widely used by experimental economists to assess effort (Gill and Prowse, 2012; Georganas et al., 2015; Besley and Ghatak, 2017; Brown et al., 2019). Originally provided in the laboratory environment, slider tasks consist of several sliders on a single screen that must be positioned in the middle between 0 and 100. To ensure consistency, each subject uses the same computer with the same mouse and keyboard settings. However, to limit the learning effects, each slider was displayed in a different position rather than below the others. For the master's thesis project, it was not common to use a laboratory, and thus I could not ensure that individuals would be participating with similar computer settings. As a result, inspired by Gill and Prowse's (2012) experiment, I created a mobile effort experiment. In Qualtrics, the study was set up as a survey, and 40 sliders with random numbers ranging from 0 to 100 were assigned (Figure 3). The initial position was set to 0. I chose to use random numbers instead of the middle 50 to avoid learning effects, as each slider is displayed directly after the other. Respondents were asked to position sliders based on the random number displayed above. According to the imposed rules, subjects were only eligible for payments if each slider in the row was correctly positioned. Individuals were progressing to the next slider after completing the one above by scrolling down.

iPhone owners were primarily targeted for employment, as Android operating systems would suffer from greater heterogeneity due to greater differences between models and mobile phones (ZcomTech, 2022). Between the models, the iOS operating systems have comparable touch duration (ZcomTech, 2022), resolution (GBKSOFT, 2021), and a default browser (GeeksforGeeks, 2021). These specifications were critical to ensure that the groups were similar to one another, and, thus, we could compare similar effort exertion in the between-subject design. However, using only one device may limit the internal and external validity. Since iPhones are more expensive on average than Android devices (Netter, 2020), these individuals may

[4] The payment was calculated based on the pilot study [N = 37] with an approximate time taken 3 minutes. Hence, the flat fee hourly averaged was £7.6. However, the filling of Prolific ID completion code and extended instruction likely increased the time to 4 minutes, therefore the individual average hourly payment in the original experiment was lower. The pilot study was conducted in May, 2021.

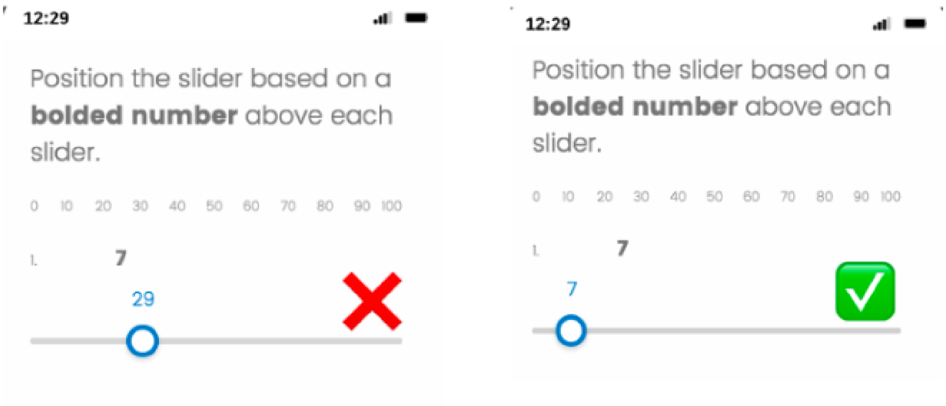
have a higher income. If individuals with higher income are more competitive, the unobserved characteristic in our study would induce upward biassed estimates.

Real-effort tasks often suffer from drawbacks. First, the cost of effort varies among subjects and is unobserved by the researcher (Gross et al., 2015; Charness et al., 2018). Second, the panel data collected from effort tasks, such as ski jumping behaviour, is noisy due to subject variation (Gill and Prowse, 2019). According to Gill and Prowse (2019), slider tasks may be perceived as advantageous from various perspectives. First, they are relatively simple and straightforward for the participants to conduct. Second, they provide nearly identical repetition compared to the math or counting exercises frequently used in effort tasks. For the same reason, they eliminate randomness in guessing. Lastly, the slider task provides a more graduated measure than the envelope-stucking task (Konow, 2000) due to the limited time frame. The slider task was originally used for repeated observations in the within-subject design, but the same benefits apply to the between-subject design.

Alternatively, the slider task has some drawbacks. First, the output of the sliders moved has no intrinsic value. Second, while there is some evidence of gender effects (Gill and Prowse, 2014). The evidence was found in the computer environment, and the results from mobile devices may differ. Gender balance was also predetermined in Prolific. Slider tasks require high concentration rather than cognitive skills. The latter could be a source of concern for the study. Once the slider was correctly positioned, a careless touch would have moved it one point, causing it to be incorrect. Sliders that were incorrectly positioned were featured (Figure 3). Respondents were asked to return and adjust the slider position. The action is time-consuming. To counter this drawback, the number of correctly moved sliders is also analysed as the outcome for the analysis.

4.2 Experimental Procedure and Hypothesis

Instructions were presented in the beginning of the tournament, explaining the rules and awarding the prizes (Appendix A). The picture of the slider (Figure 3) was also provided, but it was not interactive to ensure that participants would not be involved in any practice, which could induce learning effects prior to the tournament. They were asked to read the instructions carefully



Position the slider based on a **bolded number** above each slider.

Position the slider based on a **bolded number** above each slider.

Figure 3: An Example of the Mobile Slider Task

and thus answer the attention checks[5].

In the second stage, participants competed in a tournament. They were given a time limit and instructed to move as many sliders as possible from 0 to 100 to the required position. The tournament lasted 90 seconds. Participants were randomly assigned to one of three groups at this stage: baseline treatment (no reference point), static reference point treatment, or dynamic reference point treatment. The reference point in this experiment refers to the pilot study's highest number of moved sliders, which was 19 sliders. There was no information on a reference point provided at baseline.

The static and dynamic treatments reflect the inconclusive results from delta (distance from the line and landing location). It seems that the imposed reference points increase the distance between the reference point and the landing position, implying that the reference point motivates athletes. According to the literature, imposing competitive goals may lead to disappointment aversion (Gill and Prowse, 2012). The matter seems to be where individuals receive information about the goal on the progress line, similar to the goal-gradient-effect (Wallace et al., 2018). Individuals are more likely to achieve a goal if they perceive themselves to be closer to it. Therefore, static and dynamic treatments suggest that the reference point may affect performance depending on when the information is revealed.

Therefore, the static treatment stated the reference point before the task begins and on top of the screen (see Fig B2 in Appendix B). Gill and Prowse (2012) measured the discouragement effect

[5] Ten individuals failed the attention check, therefore they were rejected from the study and did not participate in the tournament.

discussed in the literature by using the same method of displaying reference points. Therefore, the question is whether the results will be consistent with the Gill and Prowse (2012) study or rather positive, as in the Freeman et al. (2010) experiment. The static reference point refers to the first Hypothesis:

H0: A static reference point has no effect on the individual's performance compared to no reference point.

H1: A static reference point has an effect on the individual's performance compared to no reference point.

Previous research and empirical findings from ski jumping inspired the introduction of the dynamic reference point condition. Individuals in the study were solving the slider progressively towards the reference point, similar to how ski jumpers fly towards the goal. The dynamic treatment informed them of the existence of a reference point before the task began; however, the precise position was revealed only when they were close to the 19th slider, which stated "pilot study best result" (see Fig. B3 in Appendix B). Therefore, the reference point was not immediately visible; it was only visible after some more sliders were solved and the screen was scrolled down (approx. 15 sliders depending on the phone screen size). This treatment supports the second hypothesis.

H0: A dynamic reference point has no effect on the individual's performance compared to no reference point.

H2: A dynamic reference point has an effect on the individual's performance compared to no reference point.

Lastly, after the tournament for static and dynamic treatments, the question whether the respondents remembered seeing the reference point was displayed. Following that, the study included a brief demographic questionnaire. The question of how long the current mobile phone has been in use was asked because phone usage and duration may affect performance. Device type, browser, and resolution were automatically recorded as metadata points; these variables were important for the study because differences in the environment could affect the players' performance.

It is worth noting that the experiment includes attention tests. The first ones were introduced early on when users were asked to answer questions about the game's rules. If incorrect answers were provided, the experiment was cancelled, and no earnings were received.

5 Analysis

I collected data from two sources. Experimental data have been collected with the Qualtrics survey. In addition, Prolific shares the demographics of participants; therefore, the nationality and employment status of the subjects have been matched based on the anonymous Prolific ID. The experiment had two sessions and contained a total of 172 participants. The first session was conducted on February 18, 2022, with 150 participants. The second session took place on March 4, 2022. I proceeded with two sessions, because in the first session, 22 participants needed to be rejected: 6 respondents dropped out because of a non-qualified device; 11 individuals failed an attention check; two individuals cheated by entering the study more than one time[6]; three subjects showed low effort by missing a substantial number of sliders in the row. Therefore, in order to reach the number of planned participants, the study was republished a few weeks later.

Importantly, due to the limited number of participants, I performed a two-sample means test to estimate the power of my analysis. I did it separately by device, and thus, discovered that the iOS sample size for the static treatment has a power of around 9%, whereas for the dynamic treatment, only 6% (see Table C2 in Appendix C) depending on the outcome. Furthermore, due to the higher variation in the distribution, Android has a higher power for the treatment groups (see Table C3 in Appendix C). The static condition shares a power estimate of around 22%, and the dynamic treatment has a power of around 28%[7].

Two dependent variables were explored. First, Moved [N = 151] refers to the number of total sliders moved if the number of misplaced sliders is lower than 10% of the score. A total of 19 subjects placed one or two sliders incorrectly, but still qualified for the 10% threshold. This is the main outcome variable based on the pre-registration. Second, I explored the number of correctly moved sliders. This score [N = 132] was only counted, if all sliders in the row

[6] Entering the study more than one time induces learning effects. Data provided evidence that each additional participation increased the score. Therefore they have been rejected from the study.

[7] Erasmus School of Economics policy induces a budget constraint for the Master thesis experiments. Thus 150 subjects were the highest number of participants I could employ. Thus, having a smaller amount of agents had an effect on small sample power.

were positioned correctly. Hence, if any slider in the row was missed or incorrectly placed, the score in the data set was coded as missing. It is necessary to analyse the variables separately, as placing the slider correctly was time-consuming. Thus, agents avoiding correction might reflect some risk-seeking behaviour.

The independent variables refer to baseline, static, and dynamic treatments. Each treatment variable in the study was coded as dummy. The baseline was 1 if subjects took part in the first treatment, and 0 if static or dynamic. Static and dynamic treatments were compared relative to baseline outcomes. Static versus dynamic treatment comparisons were also explored; however, this was not pre-registered and entered as an explanatory analysis. The random allocation was successful, and approximately 30 subjects per condition were assigned to the iOS operating system and 20 to the Android.

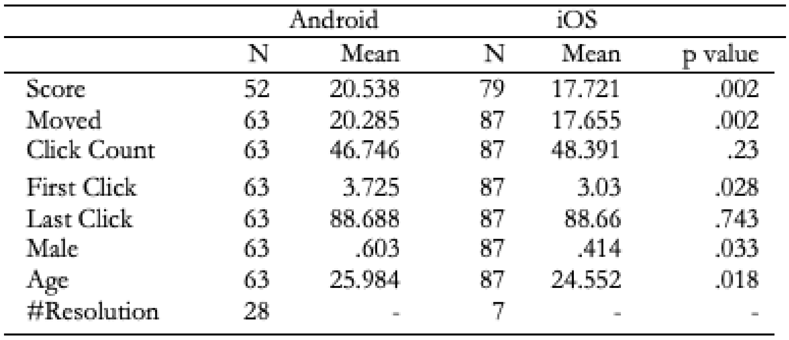
In line with pre-registration, the number of moved sliders qualifying for the 10% threshold was the main outcome. Treatments were compared relative to the baseline. The analysis needed to be carried out, aggregating both device outcomes. However, in the next section, I further explain why the decision to improve the analysis was taken.

5.1 Devices

It was necessary to ensure that individuals used similar devices to obtain homogeneous responses in the mobile slider task tournament. The study was conducted in the United States because a higher proportion of the population uses the iOS operating system. 79 respondents used the iOS operating system in the experiment. Nevertheless, a substantial 40% proportion of respondents used the Android [N = 63] operating system. One individual used a smartphone with the Windows NT operating system. It has been eliminated from the analysis.

In line with expectations, the respondents differ in terms of correctly moved sliders by device (TABLE 2). To compare the significant difference between the two samples, the Mann-Whitney test was implemented. On average, individuals holding an iPhone correctly moved 17.72 sliders, whereas Android respondents scored 20.54 (N_{iOS} = 79, $N_{Android}$ = 52; p-value = 0.002). One Android respondent correctly moved all 40 sliders, as some Android phones have touch pens, this plausibly explains this outlier. The differences between the device characteristics were more concerning. For example, respondents using Android had 28 different resolutions registered, whereas iOS had 7. The Mann-Whitney test provided a statistical difference be-

Table 2: Comparison between Android and iOS respondents



	Android		iOS		
	N	Mean	N	Mean	p value
Score	52	20.538	79	17.721	.002
Moved	63	20.285	87	17.655	.002
Click Count	63	46.746	87	48.391	.23
First Click	63	3.725	87	3.03	.028
Last Click	63	88.688	87	88.66	.743
Male	63	.603	87	.414	.033
Age	63	25.984	87	24.552	.018
#Resolution	28	-	7	-	-

Note: Resolution captures the number of the different types of screen sizes.

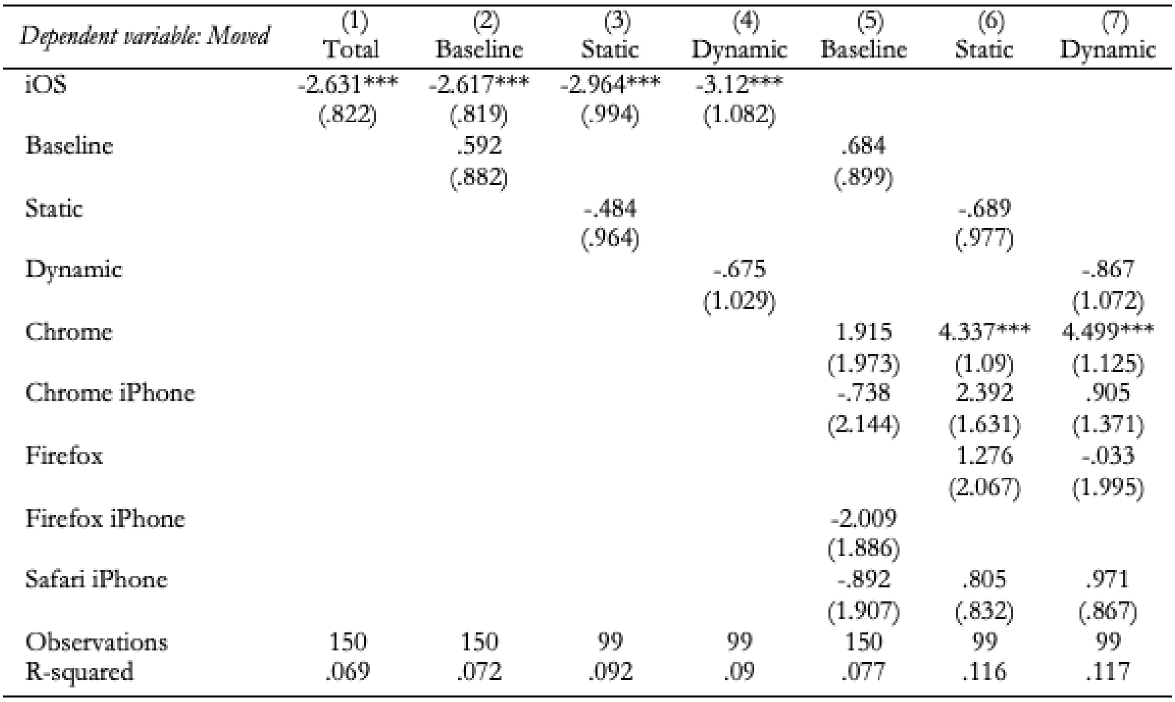
tween the time of the first click response ($N_{Android}$ = 63, N_{iOS} = 87, p-value = 0.028). One could argue that the difference could be related to the selection bias, this cannot be neglected, especially due to the significant difference between the age ($N_{Android}$ = 63, N_{iOS} = 87, p-value = 0.033) and gender ($N_{Android}$ = 63, N_{iOS} = 87, p-value = 0.018).

Pre-registration states that an OLS regression (Appendix D) is implemented (Table 3). The effect of treatment on performance was estimated while controlling for the device and browser (Columns 5 - 7). The estimates for devices are omitted due to time-invariant characteristics between browsers and devices. In order to obtain estimates, the browser dummies were excluded. The estimates for all three treatments suggest that having an iOS compared to an Android decreases the correctly moved sliders from 2.6 points (Column 2) to 3.33 points (Column 4), depending on the treatment at a 1% significance level. Lastly, the total effect of the iOS compared to Android is also highly significant at the 1% significance level. Thus, the former analysis leads to some intermediate conclusions:

Result 1: *There is a significant difference between devices. Thus, the average effect between treatments cannot be observed as it is hindered by heterogeneous device effects.*

Considering the low power of the study, these results were relatively strong, thus providing us with evidence that there is a significant difference between iOS and Android agents on the exerted effort. Suggested differences between the Android users hinder the treatment evaluation; therefore, descriptive statistics were presented separately by the operating system. The analysis was carried out accordingly with a focus on iPhone users. Noteworthy, this was not pre-registered; nevertheless, due to observed differences, this was presumed necessary.

Table 3: The Effect of the Device on the Number of Moved Slider



Dependent variable: Moved	(1) Total	(2) Baseline	(3) Static	(4) Dynamic	(5) Baseline	(6) Static	(7) Dynamic
iOS	-2.631***	-2.617***	-2.964***	-3.12***			
	(.822)	(.819)	(.994)	(1.082)			
Baseline		.592			.684		
		(.882)			(.899)		
Static			-.484			-.689	
			(.964)			(.977)	
Dynamic				-.675			-.867
				(1.029)			(1.072)
Chrome					1.915	4.337***	4.499***
					(1.973)	(1.09)	(1.125)
Chrome iPhone					-.738	2.392	.905
					(2.144)	(1.631)	(1.371)
Firefox						1.276	-.033
						(2.067)	(1.995)
Firefox iPhone					-2.009		
					(1.886)		
Safari iPhone					-.892	.805	.971
					(1.907)	(.832)	(.867)
Observations	150	150	99	99	150	99	99
R-squared	.069	.072	.092	.09	.077	.116	.117

Note. The iOS estimates were omitted in Column 5-7, because it is time-invariant with the browser dummies. Standard errors are in parentheses *** p<.01, ** p<.05, * p<.1

5.2 Descriptive Statistics

In upfront analyses, it was important to review the balance in demographic characteristics and device settings between the treatment conditions pooled by operating systems (see Table E1 Appendix E). The average age of our respondents was 25 years old. The iOS treatments shared the age of 25 between all three groups, whereas the Android users between all three groups were around 26. The overall sample was equally distributed with regard to gender due to pre-assignment in Prolific. However, some differences across the treatments were observed. For iOS, 56% of the respondents were male in the baseline condition, whereas for the Static and Dynamic groups, it was around 35% of the male respondents. For Android, 48% of the respondents were male in the baseline treatment, in static - 64% and in dynamic - 70%[8]. The study was published in the United States; therefore, it is not surprising that above 75% of the responses per condition have United States nationality. The variable was presented as a dummy variable; the United States is 1, and others nationalities have been coded as 0, as the distribu-

[8] Unequal distribution for Android was slightly concerning. This again provides us with evidence that Android results should be analysed with caution.

tion was relatively widely spread. Education variables resemble random distributions between the treatments. Most of the respondents held bachelor's degrees, and slightly fewer individuals had some college but no degree. Lastly, the information regarding an individual's employment status was obtained from Prolific. Most of the subjects were fully employed. Nevertheless, a substantial amount of individual information was missing; therefore, this variable might suffer from attrition bias and thus should be treated carefully. This leads to some intermediate conclusions:

Result 2: *Demographic characteristics are well balanced between the treatments.*

Phone behaviour characteristics were presented by the treatments (see Table E1 Appendix E). The first click appeared the fastest in the baseline condition. This might be related to the fact that the task was put in only one sentence. For static and dynamic treatments, additional reference point information were provided[9]. However, based on the Kruskal-Wallis equality-of-populations rank test, the difference between the treatments is not significant for both iOS ($N_{baseline}$ = 27, N_{static} = 29, $N_{dynamic}$ = 31, p-value = 0.6786) and Android ($N_{baseline}$ = 21, N_{static} = 22, $N_{dynamic}$ = 20, p-value = 0.7625). The last click distribution is equal between the operating systems and across the treatments. Click count differed slightly between the treatments. The baseline treatments generated the most clicks, which might be expected for iOS, because in this condition, the least number of missed sliders was observed, suggesting that individuals in the baseline treatment were more careful (Table 4). However, it could be related to the number of moved sliders for Android (Table 4), as the more sliders one positions, the more clicks it requires. Progressively lower numbers of clicks was witnessed for the static and dynamic treatments for both Android and iOS, which was in line with decreasing performance in the score. Kruskal-Wallis equality-of-populations rank test could not reject the null hypothesis; thus, there was no significant difference between the treatments. Nevertheless, the p-values were closer to 10% for Android ($N_{baseline}$ = 21, N_{static} = 22, $N_{dynamic}$ = 20, p-value = 0.1292), but not for iOS ($N_{baseline}$ = 27, N_{static} = 29, $N_{dynamic}$ = 31, p-value = 0.2942). Hence, this leads to some intermediate conclusions about phone characteristics:

Result 3: *Phone characteristics are balanced between the treatments but not across devices.*

Lastly, data on the current phone usage was obtained. Subjects from iOS had been using the

[9] The screen was adjusted accordingly. Even though the baseline had only one sentence, empty lines were added below, in order to match the task starting point to static and dynamic treatments.

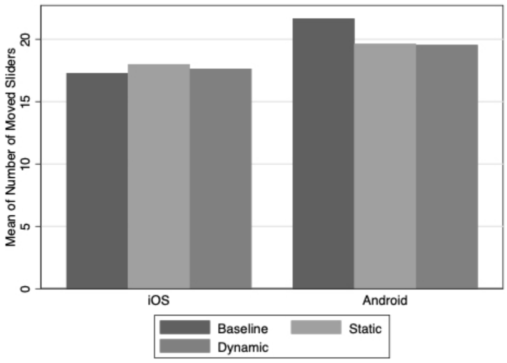


Figure 4: Mean Comparison between treatments split by devices

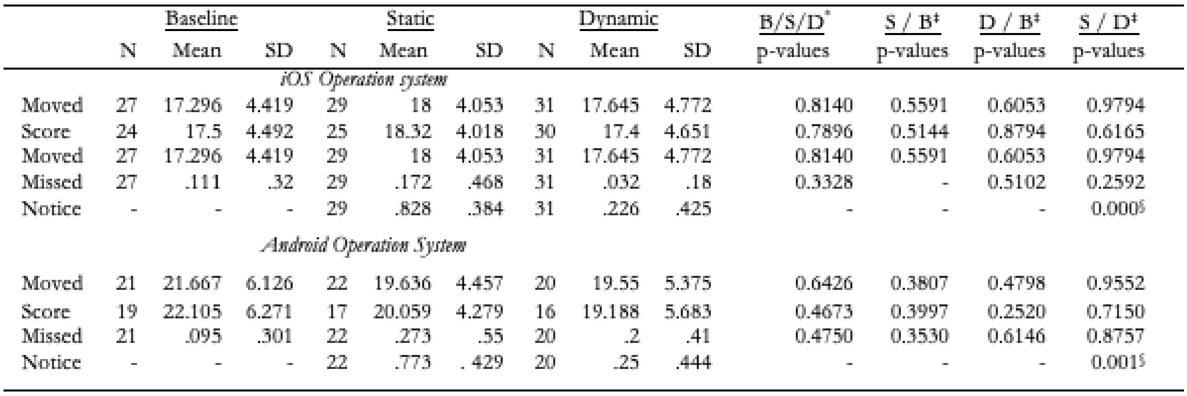
phone for about 2.5 years. Android respondents differ significantly across treatments, ranging from 3.2 years in the baseline treatment to 2.2 years in the dynamic treatment. The balance between the years of phone usage was important because it could reflect two obstacles. First, the experience of using the phone. Second, longer phone durability may result in response time issues. Typically, outdated phones become less operable over time (BBC, 2020). In the case of the study, this would hinder the results because the effort would be combined with the response time of a smartphone. Therefore, this provides me with additional evidence that we should proceed with caution when dealing with Android responses and instead focus on iOS.

5.3 Tests on Reference points

For results, the main outcomes were presented separated by the device (Figure 4). This is different to pre-registration due to explained device diversity in subsection 5.1. Correspondent to the pre-registration Kruskal-Wallis test for number of moved sliders were performed. The Mann-Whitney U test was used to compare static and dynamic treatments relative to baseline. Furthermore, i) the outcome variables score, missed, and noticed were presented as supporting evidence, and ii) a comparison of static and dynamic treatments was entered as explanatory analysis. The former and latter were not pre-registered.

Table 4 presents the results of the collected main outcomes and the number of moved sliders. The baseline condition has the lowest mean of 17.3 for iOS, dynamic is a bit higher with a mean of 17.65, and static has the highest moved number of sliders with a value of 18. The null hypothesis cannot be rejected when carrying out the Kruskal-Wallis test on the number of moved sliders. No significant differences were found in medians between the treatments of

Table 4: The Effect of the Device on the Number of Moved Slider



	Baseline			Static			Dynamic			B/S/D*	S / B‡	D / B‡	S / D‡
	N	Mean	SD	N	Mean	SD	N	Mean	SD	p-values	p-values	p-values	p-values
					iOS Operation system								
Moved	27	17.296	4.419	29	18	4.053	31	17.645	4.772	0.8140	0.5591	0.6053	0.9794
Score	24	17.5	4.492	25	18.32	4.018	30	17.4	4.651	0.7896	0.5144	0.8794	0.6165
Moved	27	17.296	4.419	29	18	4.053	31	17.645	4.772	0.8140	0.5591	0.6053	0.9794
Missed	27	.111	.32	29	.172	.468	31	.032	.18	0.3328	-	0.5102	0.2592
Notice	-	-	-	29	.828	.384	31	.226	.425	-	-	-	0.000§
					Android Operation System								
Moved	21	21.667	6.126	22	19.636	4.457	20	19.55	5.375	0.6426	0.3807	0.4798	0.9552
Score	19	22.105	6.271	17	20.059	4.279	16	19.188	5.683	0.4673	0.3997	0.2520	0.7150
Missed	21	.095	.301	22	.273	.55	20	.2	.41	0.4750	0.3530	0.6146	0.8757
Notice	-	-	-	22	.773	. 429	20	.25	.444	-	-	-	0.001§

Note: SD denotes standard deviation. B denotes the baseline treatment condition. S denotes the static reference point condition. D denotes the dynamic reference point condition. *Kruskal–Wallis one-way analysis of variance was performed for comparison between all treatments; ‡Mann–Whitney U test was employed for pair comparison. § Notice was coded as a dummy, therefore Fisher's exact test p-values was reported.

iOS respondents ($N_{baseline}$ = 27, N_{static} = 29, $N_{dynamic}$ = 31, p-value = 0.8140). Furthermore, the Mann-Whitney U test results show no evidence of the difference between the static and baseline groups ($N_{baseline}$ = 27, N_{static} = 29, p-value = 0.5591). The null hypothesis stating that the dynamic and baseline treatment are similar was not rejected (Mann-Whitney U test, $N_{baseline}$ = 27, $N_{dynamic}$ = 31, p-value = 0.6053). P-values > 0.10 for iOS suggest that raised null hypothesis 1 and null hypothesis 2 were not to be rejected; however, the low power of the experiment should be taken into account. Next, the exploration between the dynamic and static conditions was carried out. The comparison between the static and dynamic reference point treatments did not lead to any significant differences. For the latter, no null hypotheses were rejected, so there was no difference in the number of moved sliders between the groups. (Mann-Whitney U test, N_{static} = 29, $N_{dynamic}$ = 31, p-value = 0.9794).

The score was different from the number of moved sliders due to presumed careful behaviour. A mean score of 17.5 was achieved for the baseline group for iOS. Approximately 18.32 sliders were correctly positioned for the static group. The dynamic treatment scored a similar mean to the baseline of 17.4. The similarities between baseline and dynamic treatments were not surprising. For both treatments, reference points were identified upfront. Even though the descriptive statistics presented noticeable differences for the static treatment, the non-parametric tests imply the opposite. After pooling the score medians by the treatment and performing the Kruskal-Wallis test, we cannot reject the null hypothesis ($N_{baseline}$ = 24, N_{static} = 25, $N_{dynamic}$ =

30, p-value = 0.7896). Hence, the differences between the baseline, static and dynamic treatment score medians were not statistically significant. Comparing the static treatment against the baseline, no support for the observed difference was found, and thus, the null hypothesis was not rejected ($N_{baseline}$ = 24, N_{static} = 25, p-value = 0.5144). Predictably, no systematic difference could be observed between dynamic and baseline treatment (Mann-Whitney U test, $N_{baseline}$ = 24, $N_{dynamic}$ = 30, p-value = 0.8794). The null hypothesis was not rejected when comparing static and dynamic; therefore, there is no difference between the conditions (Mann-Whitney U test, N_{static} = 25, $N_{dynamic}$ = 30, p-value = 0.6165). Based on the score and the number of moved sliders outcome, the intermediate conclusion leads to:

Result 4: *No significant differences in the main outcomes could be observed between the treatments. Nevertheless, performance is highest for the static treatment group. Baseline and dynamic performance are alike.*

On Android, the number of moved sliders and score patterns differ. An average number of 21.67 moved sliders was achieved for baseline, a mean of 19.64 was registered for static and 19.55 for dynamic treatment. The patterns were different from those on iOS due to the outlier in the baseline condition[10]. However, similar results to those on iOS of the Kruskal-Wallis test were obtained for the number of moved sliders ($N_{baseline}$ = 21, N_{static} = 22, $N_{dynamic}$ = 20, p-value = 0.4673). Supportingly, no evidence is found in the pair comparison between static and baseline ($N_{baseline}$ = 22, N_{static} = 20, p-value = 0.3807) and dynamic and baseline ($N_{baseline}$ = 21, $N_{dynamic}$ = 20, p-value = 0.4798). The mean score for baseline was 22.105, for static it amounted to 20.01, and for dynamic it was 19.88. Even though the baseline mean was affected by the outlier, Kruskal-Wallis tests provided the results on the rank-sum; thus, the null hypothesis cannot be rejected. ($N_{baseline}$ = 19, N_{static} = 17, $N_{dynamic}$ = 16, p-value = 0.6426).

The non-parametric tests' evidence suggests that there was no significant difference between the three conditions. A comparison between the static and baseline treatment and the static and dynamic condition supported this. Nevertheless, an average power of 10% in the experiment might be plausibly hiding the effects, and thus, this should be taken into consideration while interpreting the results. Hence, hypothesis 1 and hypothesis cannot fully be rejected.

[10] The highest mean was observed for baseline condition, which by accounting for the standard deviation was affected by the outlier score of 40. Different outcomes and the existing noise due to the high variation, again suggest that the results might be affected by other unobserved characteristics.

5.4 Tests on Noticing Reference Point

Hints of reference point effects could plausibly be found in remembering the reference point position. Static and dynamic treatments captured whether agents noticed or remembered the reference point. The question asked participants to identify the number of reference points; 41 subjects identified 19; five agents identified 18; two wrote 20; eight individuals noted 64 (which was the slider they needed to position on). Furthermore, two answers stated that they remembered seeing the reference point, but could not recall the number. This was under the dynamic treatment allocation. The mentioned identifications were coded as noticed (yes), also those which did not refer to the exact reference point of 19. One could argue that an inability to remember the exact number is not the same as noticing the reference point. However, the goal of the study was to measure the effect of the reference point on the performance, which, as discussed in the literature and the empirical example of ski jumping, might be unconscious. Thus, this especially applies to the dynamic condition. Expectantly, static treatment had the highest number of individuals noticing the reference point. 83% of the subjects from iOS and 77% agents from Android remembered the reference point. In the dynamic treatment, only 23% for iOS and 25% for Android could recall the reference point position. This is not surprising, because the reference point was displayed at the beginning, whereas for the dynamic condition, it was only at the point. Support for the existing differences was encountered from Fisher's Exact Test while comparing notice and static versus dynamic dummy variables. Since the p-value<0.001, the null hypothesis was rejected, and thus, there was a significant difference between the static and dynamic groups. The estimates are similar for both the iOS and Android operating systems. The results were robust, accounting for only those individuals who had a higher probability of encountering the reference point based on their screen resolution (Appendix F), leading to the intermediate conclusion:

Result 5: *A significantly higher number of agents noticed the reference point in the static treatment relative to the dynamic treatment.*

5.5 Regression Analysis

The results from the non-parametric tests do not support the hypothesis raised. The power of 10% in the experiment might have contributed to the obstacle. Nevertheless, rank-sum comparisons from the non-parametric tests might hinder some temporal patterns. For this reason, a simple OLS regression[11] can be performed:

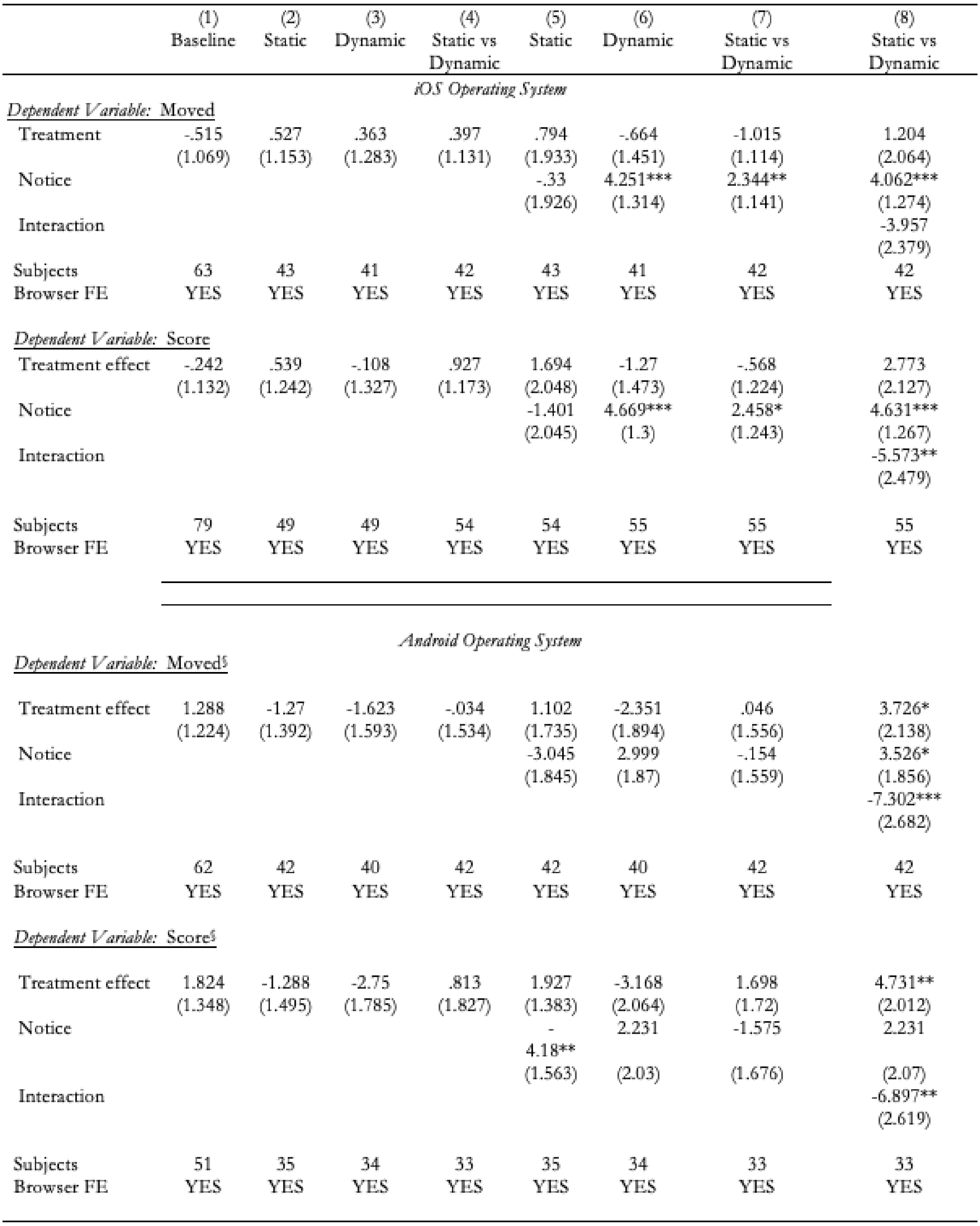
$$y_{ij} = \alpha + \beta Treatment_j + \gamma Notice_{ij} + \delta \times I(browser)_{ij} + \varepsilon_{ij} \tag{2}$$

The dependent variable indicates the score of the total number of (correctly) moved sliders for the individual *i* in the treatment *j*, and the estimate identifies the effect of the treatment condition. The control variable notice of the reference point was included. Browser fixed effects were added. Lastly, an assumption of the error term being zero was expected, as due to the randomization in the experiment no correlation between dependent and independent variables inducing the omitted variable bias plausibly occurs. The OLS regression is performed separately by the operating system and is presented in Table 5. Three different equations were implemented: i) reduced form: the effect of treatment on performance; ii) equation (2): the effect of treatment while controlling for notice; iii) interaction term: the effect of treatment on performance for individuals who noticed the reference point.

Column 1 denotes the effect of the baseline condition on the number of moved sliders. We can see that the effect is negative for iOS, suggesting that being assigned to the tournament without reference points compared to other tournaments decreases agents' performance. The same results can also be observed for the score performance. However, the effect is highly insignificant in both cases. Static treatment (Column 2) for iOS leads to positive estimates in the number of moved sliders and score, implying that perceiving information about the reference point might be encouraging. Nevertheless, the effects here are again not significant. Dynamic reference point effects are positive for the number of moved sliders and the opposite for the score; hence, the effect is close to zero. This suggests that the baseline condition estimates were most likely induced by positive coefficients in the static treatment. In Column 4, a positive estimate was observed for the static treatment compared to the dynamic condition in moved sliders. The estimates are consistent with the performance score, yet, not significant. The estimated positive reference point coefficient was similar to the Freeman et al. (2010) findings, providing evidence that information about the competitors' positions increases the number of solved mazes. On the other hand, the positive static outcome relative to baseline might point to goal-gradient-effects (Wallace and Etkin, 2017). Hence, the group with a clear goal was focused to the end goal, and thus, the baseline treatment agents referred to 0 as their reference

[11] This regression was not pre-registered. The pre-registered analysis was carried out in Table 3.

Table 5: Regression Results by Operating Systems



	(1) Baseline	(2) Static	(3) Dynamic	(4) Static vs Dynamic	(5) Static	(6) Dynamic	(7) Static vs Dynamic	(8) Static vs Dynamic
					iOS Operating System			
Dependent Variable: Moved								
Treatment	-.515	.527	.363	.397	.794	-.664	-1.015	1.204
	(1.069)	(1.153)	(1.283)	(1.131)	(1.933)	(1.451)	(1.114)	(2.064)
Notice					-.33	4.251***	2.344**	4.062***
					(1.926)	(1.314)	(1.141)	(1.274)
Interaction								-3.957
								(2.379)
Subjects	63	43	41	42	43	41	42	42
Browser FE	YES	YES	YES	YES	YES	YES	YES	YES
Dependent Variable: Score								
Treatment effect	-.242	.539	-.108	.927	1.694	-1.27	-.568	2.773
	(1.132)	(1.242)	(1.327)	(1.173)	(2.048)	(1.473)	(1.224)	(2.127)
Notice					-1.401	4.669***	2.458*	4.631***
					(2.045)	(1.3)	(1.243)	(1.267)
Interaction								-5.573**
								(2.479)
Subjects	79	49	49	54	54	55	55	55
Browser FE	YES	YES	YES	YES	YES	YES	YES	YES
					Android Operating System			
Dependent Variable: Moved§								
Treatment effect	1.288	-1.27	-1.623	-.034	1.102	-2.351	.046	3.726*
	(1.224)	(1.392)	(1.593)	(1.534)	(1.735)	(1.894)	(1.556)	(2.138)
Notice					-3.045	2.999	-.154	3.526*
					(1.845)	(1.87)	(1.559)	(1.856)
Interaction								-7.302***
								(2.682)
Subjects	62	42	40	42	42	40	42	42
Browser FE	YES	YES	YES	YES	YES	YES	YES	YES
Dependent Variable: Score§								
Treatment effect	1.824	-1.288	-2.75	.813	1.927	-3.168	1.698	4.731**
	(1.348)	(1.495)	(1.785)	(1.827)	(1.383)	(2.064)	(1.72)	(2.012)
Notice					- 4.18**	2.231	-1.575	2.231
					(1.563)	(2.03)	(1.676)	(2.07)
Interaction								-6.897**
								(2.619)
Subjects	51	35	34	33	35	34	33	33
Browser FE	YES	YES	YES	YES	YES	YES	YES	YES

Note: Browser FE denotes browser fixed effects. §Regressions employed for the Android Operating system excluded the outlier number of 40 moved sliders and score 40. Standard errors are in parentheses: *** p<.01, ** p<.05, * p<.1

point. This leads to some intermediate conclusions:

Result 6: *Inducing a reference point in static and dynamic treatments relative to no reference increases the agent's performance. Results were consistent, but not significant.*

Columns 5 to 7 show control variables - whether subjects notice that the reference point was added. The static reference point relative to the baseline still provides us with consistent positive estimates for both dependent variables on the iOS operating system. Interestingly, noticing the reference point negatively affects the performance, but it is still insignificant. Consistent negative coefficients between both dependent variables suggest that agents' awareness of the existing reference point discourages goal persuadability. Dynamic reference point estimates slightly differ from the previous case. Now, both dependent variables lead to negative effects, suggesting that the dynamic treatment might also discourage the individuals. However, noticing reference point estimates imply that noticing the reference point increases the number of moved sliders by 4.25 points at a 1% significance level, ceteris paribus. The effect of noticing the reference point was consistent and led to higher estimates for scores than for moved sliders, increasing by 4.67 points at a 1% significance level, ceteris paribus. Noteworthy, it is also an economically significant effect, as noticing the reference point relatively increases the number of moved sliders by 24% (Mean 17.645) and the score by 26% (Mean 17.4). The opposite direction of the effect between static and dynamic control variables proposes that the reference point information timing affects the performance. It might be noted that , estimates of remembering the reference point here should be treated with caution: i) the baseline was not provided with a reference point; thus, observations for the baseline notice variable refer to zero; ii) in the dynamic treatment, agents were more likely to encounter reference points once more sliders had been moved and the position was closer to the reference point. This may induce overestimated coefficients. The former discussion leads to some intermediate conclusions:

Result 7: *Noticing the reference point in the static treatment leads to negative consistent but insignificant performance, relative to baseline. The results are positive relative to the dynamic treatment.*

Lastly, the effect between the static and dynamic reference points was calculated in Column 7. In Column 8, the interaction term between noticing the reference point and the static versus dynamic dummy was added. Once the control variable for remembering the point was introduced (Column 7, see iOS), the effect of the static compared to the dynamic reference point became negative and consistent between both dependent variables, yet, still insignificant. That might suggest that inducing the reference point upfront, compared to closer to the goal, limits agents' ability to exert effort; hence, it acts as a discouragement effect. Remembering the reference point compared to not remembering, increases the number of moved sliders by 2.34 points at a 5% significance level, ceteris paribus. This also had an economically significant effect, as the relative increase in the number of moved sliders is around 13% (mean 18) and the score - 13.4% (mean 18.32). In Column 8 (see iOS), the interaction term was included, allowing to interpret the effect of noticing the reference point and being assigned to a static treatment on the performance. The interaction term for moved sliders is also negative, leading to a 3.3 point decrease, ceteris paribus; however, the p-value is equal to 0.121. The interaction term provided significant results at a 5% significance level for the number of moved sliders and implies that being in a static treatment relative to dynamic and remembering the reference point decreases the performance by 5.57 points, ceteris paribus.

Similar to iOS, none of the treatment effects in Columns 1 to 4 have an effect on the performance of Android. However, the estimates were substantially higher than for iOS by more than one moved slider. The static (Column 2) and dynamic (Column 3) reference points relative to the baseline treatment for both dependent variables provide negative estimates, suggesting that the reference point discouraged agents to excel their optimal effort in the tournament. Furthermore, the effect of the static compared to the dynamic treatment was inconsistent between dependent variables, and thus, the effect was close to 0. In columns 5 to 6, the static and dynamic relative to the baseline conditions were estimated while controlling for remembering the reference point. Significant results were only observed for static reference points at a 5% significance level. The sign is also negative, which was in line with disappointment aversion due to stated information upfront of the tournament. Estimates for dynamic treatment opposite to iOS were negative and higher than for static groups. Consequently, the static versus dynamic conditions (Column 7) presented low estimates and, thus, were more likely to be close to zero. Lastly, estimates are significant and substantially high once the interaction term was introduced for measuring the effect of the static treatment relative to the dynamic treatment. The effect of the static treatment and remembering the reference point decreases the number of moved sliders by 7.3 points at a 1% significance level and the score by 6.9 points at a 5% significance level, ceteris paribus[12]. This leads to some intermediate conclusions:

[12] The robustness of the results was explored in the Appendix F. The robustness checks suggest that there is no effect, however, the estimates are consistent in sign. Thus, the effect might be hindered by low power.

Result 8: *The presence of a significant negative interaction term between static versus dynamic treatments and noticing the reference point suggests that agents experience disappointment aversions.*

6 Discussion and Limitations

Evidence suggests that Hypothesis 1 is only weakly supported. According to the hypothesis, there was an effect when compared to the baseline scenario. Based on non-parametric tests, the null hypothesis was not rejected for static and dynamic reference points relative to the baseline. Nonetheless, some weak positive effects for static reference points relative to the baseline from OLS regression were observed (Table 5). Support for Hypothesis 1 could be weakly assumed due to consistent estimates in ostensibly homogeneous groups and the low power of the estimates. Following that, no effect of the dynamic reference point relative to the baseline was found due to high p-values from the non-parametric test and inconsistencies in estimates. Therefore, we cannot reject the null hypothesis in Hypothesis 2.

Explanatory analysis on the differences between static and dynamic reference points provided evidence of the discouragement effect for Hypothesis 1. The consistency between the dependent variables on iOS suggests that a static reference point relative to a dynamic one provided a negative effect on the performance, ceteris paribus. This is contrary to the findings relative to the baseline. Furthermore, the interaction term between the static and dynamic treatment and noticing the dummies produced highly significant results, suggesting that knowing about goal-as-a-reference upfront negatively affects performance and discourages individuals from performing optimally in rank-order tournaments. The results are in line with the literature on visibility (Cheema and Bagchi, 2011) and very similar to the finding of Gill and Prowse (2012) on disappointment aversion – knowing about the reference point upfront causes agents to have worse performance. The opposite signs between the static condition relative to the baseline and the dynamic condition are linked to Wallace and Etkin's (2017) findings on goal specificity. Specific versus non-specific goals lead to ambiguity aversions, and thus individuals value their performance on how far they moved from the starting point, applying to the baseline condition. Non-specific goals are applied to baseline and partially to dynamic conditions. Based on the observed results, one could conclude that agents entered with the "do-your-best" strategy in the dynamic treatment, but the displayed reference points in the progress changed their strategy to “beat-19”. Negative effects on the interaction terms support the latter interpretation. This indirectly provides some evidence on Hypothesis 2, but more research is still required.

The experimental study suffers from some limitations, which should be discussed. To drive causal inference conclusions, experiments must ensure that the only changing condition across the treatments was the interest variable. The environment in each condition needs to be similar to capture the treatment effect. This was even more crucial for real-effort tasks because the study sought to evaluate exerted performance, assuming that each individual was provided with the same tool to participate in the study. This could be ensured in the onsite laboratory experiment by using the same computer setting as Gill and Prowse (2012). The same phone devices were examined in this experiment to increase the probability of homogeneity. Nevertheless, this did not ensure that the latter was achieved. Following this, analysis results from only one device induce selection bias in the experiment. Due to income, social acceptance, and other factors, a limited amount of randomness can be assumed for phone purchases. For example, social acceptance positively correlates with performance (Wentzel et al., 2021), and plausibly reinforces reference point achievement, thus distorting the results and inducing upward biased estimates.

Recruiting agents via online platforms could impose some biases. First, they select the study themselves, much like the devices that registered people to Prolific. However, this is not different from the laboratory experiment onsite. Following this, agents can have some opportunities to choose the study topics depending on the short study descriptions. Thus, individuals who are more interested in a certain topic might be more likely to select themselves (Prolific, 2018). Nevertheless, Prolific randomly distributes published studies to the respondents, and thus the “first-come, first-served” rule applies (Prolific, 2018). The researcher’s responsibility is to ensure that the presented description remains neutral[13]. Second, Prolific demographics suggest that the platform suffers from the “WEIRD bias” (Prolific, 2018). According to Prolific, they have a bias toward women, young people, and people with a high level of education. Hence, the results may have limited external validity. Third, there is some rapid-response bias due to the first-come, first-served rules. (Prolific, 2018) This was less of an issue in my experiment because the study was relatively short and required a quick response to win the prize. Furthermore, the attention checks for the experiment rules were implemented early on, and unqualified

[13] The description from Prolific can be seen in Appendix G.

agents were immediately rejected from the study. To summarise, Prolific may have some drawbacks that we should be aware of. However, the platform is well-liked among researchers (Palan and Schitter, 2018). Using Prolific helps to avoid any researcher-imposed demand bias because the respondents have no connection to the study owners.

According to the literature, the goal-gradient effect between specified and unclear goals has an effect on the results (Wallace and Etkin, 2018). Agents imposed with the goal of "do your best" rely on how far they have moved from the start, whereas individuals with clear goals, such as "the best results is 19 sliders," seek to achieve the 19. This might be a minor obstacle in the current experiment. Supposing that agents in the baseline treatment behaved according to the experimental findings from Wallace and Etkin (2018) and exerted more effort in the beginning, we might be overestimating the results because the baseline group score distribution is influenced to the left, while the specified goal treatment score distribution moves to the right. I could have programmed the timing function on each slider moved to measure it. This would have helped me account for the time each individual takes to move one slider. Data on the number of sliders moved by time would have gradually informed us about the exerted effort. This should be taken into consideration for future research because the baseline treatment might not be the right counterfactual if the framing relies on "do your best."

To capture whether subjects were aware of the reference point, the question of noticing the reference point was added. This points to a few limitations while interpreting the results. First, the baseline did not have any reference points; thus, we are overestimating the results found for static and dynamic reference points by coding them as zero. This was indeed questionable if this should have been done; however, the research seeks to understand the effect of the reference point on performance. If no reference point is exposed, this assumes that no reference points could be relied on. For future research, an improvement could be made by adding a question to the baseline condition to determine whether individuals had imposed any personal reference points to make the comparison more equal. Second, the comparison of noticing the reference points between the static and dynamic conditions might be overestimated. This is because agents in the dynamic treatment could notice the reference point only if they achieved a certain number of sliders, consequently imposing a higher score. A robustness check on viewing the reference point based on the screen size resolution raises concerns about a significant effect for interaction terms. Nonetheless, due to the experiment's low power, this is inconclusive and more research should be carried out.

A rather weak support for both hypotheses could be attributed to the small sample size. Results derived from a sample of 150 respondents led to the low sample power of the experiment. This has been taken into account while analysing the results. The power decreases even more due to the analyses carried out separately by devices. Although the main effects were insignificant, this does not allow the conclusion that there was no effect of the reference point on the performance in the experiment.

Lastly, tournament characteristics have an effect on the exerted effort. One could argue that different prizes or high stakes affect the performance differently, and thus the results could not be observed in the field. Nevertheless, the flat payment of £0.38 and the prize of £1.47 lead to a substantial hourly earning of £27.75. Following that, experimental evidence suggests that increasing the prize level in the tournaments has no effect (Leuven et al., 2011; Paola et al., 2018). Individuals' performances do not significantly change in the rank-order tournaments if higher payments are provided; therefore, we could conclude that my study results should not change if a higher prize or flat-rate payments are introduced. Another tournament design characteristic, which has an effect on individual performance is tournament size (Garcia and Tor, 2009; Boudreau et al., 2016). Evidence suggests that the average performance decreases when the number of competitors increases. However, this should not be a concern in our experiment. All agents were presented with the same number of competitors in the instruction and had been paid the flat rate.

Overall, the study had some limitations the reader should be aware of. Providing the analyses separately by the device and interpreting the notice of the reference point as significant effects should be interpreted with caution. On another hand, limitations considering the Prolific respondents, tournament characteristics and small sample size were lesser concerns.

7 Conclusion

Finally, the results from the empirical analysis and the experimental setting provide some evidence that individuals act in line with the Prospect Theory presented by Kahneman and Tversky (1979) and thus violate the expected utility. Therefore, the research question could be answered that imposing a reference point has an effect on individual performance. Most interestingly, the experiment sought to understand whether the noisy results from the empirical ski jumping findings could be replicated in a laboratory experiment. The findings provide evidence that reference points lead to discouragement effects, especially for those who noticed the reference point. Nevertheless, the study suffered from low sample power and more research adjusting for limitations should be conducted.

References

[1] Allen, E. J. et al. “Reference-Dependent Preferences: Evidence from Marathon Runners”. In: *Management Science* 63.6 (2017), pp. 1657–1672. DOI: 10.1287/mnsc.2015.2417.

[2] Barberis, N. C. “Thirty Years of Prospect Theory in Economics: A Review and Assessment”. In: *Journal of Economic Perspectives* 27.1 (2013), pp. 173–196. DOI: 10.1257/jep.27.1.173.

[3] BBC News. *Apple fined for slowing down old iPhones*. british. Feb. 7, 2020. URL: https://www.bbc.com/news/technology-51413724 (visited on 04/25/2022).

[4] Bell, D. . and Bucklin, R. . “The Role of Internal Reference Points in the Category Purchase Decision”. In: *Journal of Consumer Research* 26.2 (1999), pp. 128–143. DOI: 10.1086/209555.

[5] Besley, T. . and Ghatak, M. . “Profit with Purpose? A Theory of Social Enterprise”. In: *American Economic Journal: Economic Policy* 9.3 (2017), pp. 19–58. DOI: 10.1257/pol.20150495.

[6] Bognanno, M. . “Corporate Tournaments”. In: *Journal of Labor Economics* 19.2 (2001), pp. 290–315. DOI: 10.1086/319562.

[7] Bolton, G. E. and Ockenfels, A. . “ERC: A Theory of Equity, Reciprocity, and Competition”. In: *American Economic Review* 90.1 (2000), pp. 166–193. DOI: 10.1257/aer.90.1.166.

[8] Boudreau, K. J., Lakhani, K. R., and Menietti, M. . “Performance responses to competition across skill levels in rank[U+2010]order tournaments: field evidence and implications for tournament design”. In: *The RAND Journal of Economics* 47.1 (2016), pp. 140–165. DOI: 10.1111/1756-2171.12121.

[9] Brown, A. L., Meer, J. ., and Williams, J. F. "Why Do People Volunteer? An Experimental Analysis of Preferences for Time Donations". In: *Management Science* 65.4 (2019), pp. 1455–1468. DOI: 10.1287/mnsc.2017.2951.

[10] Casas-Arce, P. . "Relative performance compensation, contests, and dynamic incentives". In: *Development and Learning in Organizations: An International Journal* 24.2 (2010). DOI: 10.1108/dlo.2010.08124bad.010.

[11] Cason, T. N., Masters, W. A., and Sheremeta, R. M. "Winner-take-all and proportional-prize contests: Theory and experimental results". In: *Journal of Economic Behavior Organization* 175 (2020), pp. 314–327. DOI: 10.1016/j.jebo.2018.01.023.

[12] Charness, G. ., Gneezy, U. ., and Henderson, A. . "Experimental methods: Measuring effort in economics experiments". In: *Journal of Economic Behavior and Organization* 149 (2018), pp. 74–87. DOI: 10.1016/j.jebo.2018.02.024.

[13] Cheema, A. . and Bagchi, R. . "The Effect of Goal Visualization on Goal Pursuit: Implications for Consumers and Managers". In: *Journal of Marketing* 75.2 (2011), pp. 109–123. DOI: 10.1509/jmkg.75.2.109.

[14] Clark, D. . et al. "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments". In: *The Review of Economics and Statistics* 102.4 (2020), pp. 648–663. DOI: 10.1162/rest_a_00864.

[15] Cryder, C. E., Loewenstein, G. ., and Seltman, H. . "Goal gradient in helping behavior". In: *Journal of Experimental Social Psychology* 49.6 (2013), pp. 1078–1083. DOI: 10.1016/j.jesp.2013.07.003.

[16] De Paola, M. ., Gioia, F. ., and Scoppa, V. . "The adverse consequences of tournaments: Evidence from a field experiment". In: *Journal of Economic Behavior and Organization* 151 (2018), pp. 1–18. DOI: 10.1016/j.jebo.2018.05.001.

[17] Dechenaux, E. ., Kovenock, D. ., and Sheremeta, R. M. "A survey of experimental research on contests, all-pay auctions and tournaments". In: *Experimental Economics* 18.4 (2014), pp. 609–669. DOI: 10.1007/s10683-014-9421-0.

[18] Delfgaauw, J. . et al. "Dynamic incentive effects of relative performance pay: A field experiment". In: *Labour Economics* 28 (2014), pp. 1–13. DOI: 10.1016/j.labeco.2014.02.003.

[19] DEV, D. . *Head to Head Comparison Between iOS vs Android.* Apr. 3, 2022. URL: https://zcom.tech/ios-vs-android.html/ (visited on 03/05/2022).

[20] Eisenkopf, G. . and Teyssier, S. . "Envy and loss aversion in tournaments". In: *Journal of Economic Psychology* 34 (2013), pp. 240–255. DOI: 10.1016/j.joep.2012.06.006.

[21] Fairburn, J. A. and Malcomson, J. M. "Performance, Promotion, and the Peter Principle". In: *The Review of Economic Studies* 68.1 (2001), pp. 45–66. DOI: 10.1111/1467-937x.00159.

[22] Fershtman, C. . and Gneezy, U. . "The tradeoff between performance and quitting in high power tournaments". In: *Journal of the European Economic Association* 9.2 (2011), pp. 318–336. DOI: 10.1111/j.1542-4774.2010.01012.x.

[23] Fiegenbaum, A. ., Hart, S. ., and Schendel, D. . "STRATEGIC REFERENCE POINT THEORY". In: *Strategic Management Journal* 17.3 (1996), pp. 219–235. DOI: 10.1002/(SICI)1097-0266(199603)17:3.

[24] FIS-ski.com. *Successful premiere: The "to-beat" line - Sports Club Flying Skier - Perm - Russia*. Jan. 6, 2014. URL: http://tramplin.perm.ru/news/eng/2014/2014_01_06_01.htm (visited on 04/10/2022).

[25] FIS-ski.com. *Walter Hofer: "Two more winters, then it's over"*. Nov. 15, 2018. URL: https://www.fis-ski.com/en/ski-jumping/ski-jumping-news-multimedia/news/2018-19/walter-hofer-two-more-winters-then-it-s-over (visited on 04/06/2022).

[26] Fishbach, A. ., Eyal, T. ., and Finkelstein, S. R. "How Positive and Negative Feedback Motivate Goal Pursuit". In: *Social and Personality Psychology Compass* 4.8 (2010), pp. 517–530. DOI: 10.1111/j.1751-9004.2010.00285.x.

[27] Freeman, R. B. and Gelber, A. M. "Prize Structure and Information in Tournaments: Experimental Evidence". In: *American Economic Journal: Applied Economics* 2.1 (2010), pp. 149–164. DOI: 10.1257/app.2.1.149.

[28] Garcia, S. M. and Tor, A. . "The N-effect: more competitors, less competition". In: *Psychological Science* 20.7 (2009), pp. 871–877. DOI: 10.1111/j.1467-9280.2009.02385.x.

[29] GBKSOFT Software Development Consulting Blog. *Mobile App Screen Dimensions Resolutions for iOS Android Design | GBKSOFT*. Oct. 7, 2021. URL: https://gbksoft.com/blog/dimensions-resolution-for-ios-and-android-app-design/ (visited on 03/05/2022).

[30] GeeksforGeeks. *Difference between iOS and Android.* Sept. 3, 2021. URL: https://www.geeksforgeeks.org/difference-between-ios-and-android/ (visited on 03/05/2022).

[31] Georganas, S. ., Tonin, M. ., and Vlassopoulos, M. . "Peer pressure and productivity: The role of observing and being observed". In: *Journal of Economic Behavior Organization* 117 (2015), pp. 223–232. DOI: 10.1016/j.jebo.2015.06.014.

[32] Gill, D. . and Prowse, V. . "A Structural Analysis of Disappointment Aversion in a Real Effort Competition". In: *American Economic Review* 102.1 (2012), pp. 469–503. DOI: 10.1257/aer.102.1.469.

[33] Gill, D. . and Prowse, V. . "Gender differences and dynamics in competition: The role of luck". In: *Quantitative Economics* 5.2 (2014), pp. 351–376. DOI: 10.3982/qe309.

[34] Gill, D. . and Prowse, V. . "Measuring costly effort using the slider task". In: *Journal of Behavioral and Experimental Finance* 21 (2019), pp. 1–9. DOI: 10.1016/j.jbef.2018.11.003.

[35] Gross, T. ., Guo, C. ., and Charness, G. . "Merit pay and wage compression with productivity differences and uncertainty". In: *Journal of Economic Behavior Organization* 117 (2015), pp. 233–247. DOI: 10.1016/j.jebo.2015.06.009.

[36] Harkin, B. . et al. "Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence." In: *Psychological Bulletin* 142.2 (2016), pp. 198–229. DOI: 10.1037/bul0000025.

[37] Heath, C. ., Larrick, R. P., and Wu, G. . "Goals as Reference Points". In: *Cognitive Psychology* 38.1 (1999), pp. 79–109. DOI: 10.1006/cogp.1998.0708.

[38] Hull, C. L. "The goal-gradient hypothesis and maze learning." In: *Psychological Review* 39.1 (1932), pp. 25–43. DOI: 10.1037/h0072640.

[39] Kahneman, D. ., Knetsch, J. L., and Thaler, R. H. "Experimental Tests of the Endowment Effect and the Coase Theorem". In: *Journal of Political Economy* 98.6 (1990), pp. 1325–1348. DOI: 10.1086/261737.

[40] Kahneman, D. . and Tversky, A. . "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2 (1979), pp. 263–291. DOI: 10.2307/1914185.

[41] Konow, J. . "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions". In: *American Economic Review* 90.4 (2000), pp. 1072–1092. DOI: 10.1257/aer.90.4.1072.

[42] Konrad, K. . A. . *Strategy and Dynamics in Contests*. Oxford, United Kingdom: Oxford University Press, 2009.

[43] Koo, M. . and Fishbach, A. . "Dynamics of self-regulation: How (un)accomplished goal actions affect motivation." In: *Motivation Science* 1.S (2014), pp. 73–90. DOI: 10.1037/2333-8113.1.s.73.

[44] Koo, M. . and Fishbach, A. . "The Small-Area Hypothesis: Effects of Progress Monitoring on Goal Adherence". In: *Journal of Consumer Research* 39.3 (2012), pp. 493–509. DOI: 10.1086/663827.

[45] Kőszegi, B. . and Rabin, M. . "Reference-Dependent Risk Attitudes". In: *American Economic Review* 97.4 (2007), pp. 1047–1073. DOI: 10.1257/aer.97.4.1047.

[46] Leuven, E. . et al. "Incentives versus Sorting in Tournaments: Evidence from a Field Experiment". In: *Journal of Labor Economics* 29.3 (2011), pp. 637–658. DOI: 10.1086/659345.

[47] Levitt, S. D. and List, J. A. "Homo economicus evolves". In: *Science* 319.5865 (2008), pp. 909–910. DOI: 10.1126/science.1153640.

[48] Lim, W. ., Matros, A. ., and Turocy, T. L. "Bounded rationality and group size in Tullock contests: Experimental evidence". In: *Journal of Economic Behavior Organization* 99 (2014), pp. 155–167. DOI: 10.1016/j.jebo.2013.12.010.

[49] List, J. A. "Does Market Experience Eliminate Market Anomalies?" In: *The Quarterly Journal of Economics* 118.1 (2003), pp. 41–71. DOI: 10.1162/00335530360535144.

[50] List, J. A. "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace". In: *Econometrica* 72.2 (2004), pp. 615–625. DOI: 10.1111/j.1468-0262.2004.00502.x.

[51] Ludwig, S. . and Lünser, G. K. "Observing your competitor – The role of effort information in two-stage tournaments". In: *Journal of Economic Psychology* 33.1 (2012), pp. 166–182. DOI: 10.1016/j.joep.2011.09.011.

[52] Netter, S. . *iPhone vs Android: Which is better for your finances?* Sept. 2, 2020. URL: https://www.varomoney.com/money/android-vs-iphone/.

[53] Niederle, M. . and Vesterlund, L. . "Gender and Competition". In: *Annual Review of Economics* 3.1 (2011), pp. 601–630. DOI: 10.1146/annurev-economics-111809-125122.

[54] Nunes, J. . and Drèze, X. . “The Endowed Progress Effect: How Artificial Advancement Increases Effort”. In: *Journal of Consumer Research* 32.4 (2006), pp. 504–512. DOI: 10.1086/500480.

[55] Palan, S. . and Schitter, C. . “Prolific.ac—A subject pool for online experiments”. In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27. DOI: 10.1016/j.jbef.2017.12.004.

[56] Perelman, R. . *SKI JUMPING: Kobayashi equals all-time World Cup record of six consecutive wins*. Jan. 14, 2019. URL: http://www.thesportsexaminer.com/ski-jumping-kobayashi-equals-all-time-world-cup-record-of-six-consecutive-wins/ (visited on 04/10/2022).

[57] Prolific Team. *What are the advantages and limitations of an online sample?* british. Sept. 18, 2018. URL: https://researcher-help.prolific.co/hc/en-gb/articles/360009501473-What-are-the-advantages-and-limitations-of-an-online-sample- (visited on 03/28/2022).

[58] Shoham, A. . and Fiegenbaum, A. . “Extending the Competitive Marketing Strategy Paradigm: The Role of Strategic Reference Points Theory”. In: *Journal of the Academy of Marketing Science* 27.4 (1999), pp. 442–454. DOI: 10.1177/0092070399274004.

[59] Statista. *iPhone users as share of smartphone users in the United States 2014-2021*. Mar. 31, 2021. URL: https://www.statista.com/statistics/236550/percentage-of-us-population-that-own-a-iphone-smartphone/#:%7E:text=How%20many%20people%20have%20iPhones,users%20in%20the%20United%20States..

[60] *THE INTERNATIONAL SKI COMPETITION RULES (ICR)*. SKI JUMPING. Vol. BOOK III. International Ski Federation FIS. URL: https://assets.fis-ski.com/image/upload/v1639755981/fis-prod/assets/ICR_Ski_Jumping_2022_clean.pdf.

[61] Tversky, A. . and Kahneman, D. . “Advances in prospect theory: Cumulative representation of uncertainty”. In: *Journal of Risk and Uncertainty* 5.4 (1992), pp. 297–323. DOI: 10.1007/bf00122574.

[62] Virmavirta, M. . and Kivekäs, J. . “The effect of wind on jumping distance in ski jumping – fairness assessed”. In: *Sports Biomechanics* 11.3 (2012), pp. 358–369. DOI: 10.1080/14763141.2011.637119.

[63] Wallace, S. G. and Etkin, J. . “How Goal Specificity Shapes Motivation: A Reference Points Perspective”. In: *Journal of Consumer Research* 44.5 (2017), pp. 1033–1051. DOI: 10.1093/jcr/ucx082.

[64] Wang, P. . et al. “Reference points in consumer choice models: A review and future research agenda”. In: *International Journal of Consumer Studies* 45.5 (2020), pp. 985–1006. DOI: 10.1111/ijcs.12637.

[65] Wentzel, K. R., Jablansky, S. ., and Scalise, N. R. “Peer social acceptance and academic achievement: A meta-analytic study.” In: *Journal of Educational Psychology* 113.1 (2021), pp. 157–180. DOI: 10.1037/edu0000468.

[66] Wright, P. M. and Kacmar, K. . “Goal Specificity as a Determinant of Goal Commitment and Goal Change”. In: *Organizational Behavior and Human Decision Processes* 59.2 (1994), pp. 242–260. DOI: 10.1006/obhd.1994.1059.

Appendices

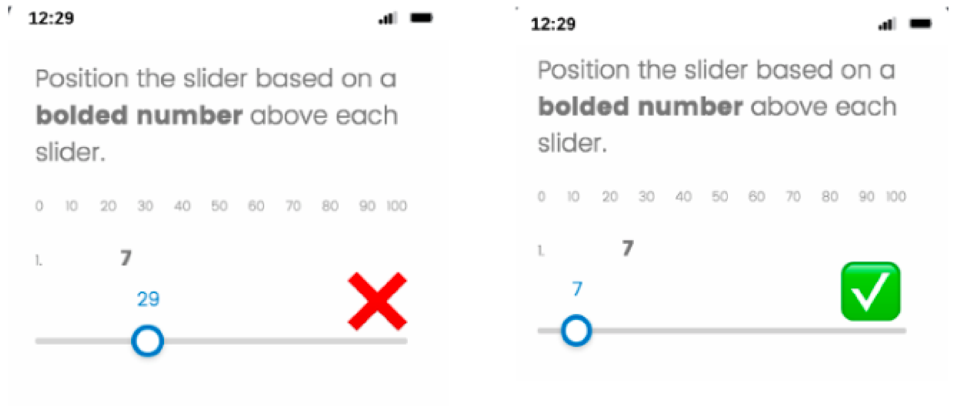
Appendix A Instructions of the experiment provided for the respondents of the study.

Instructions

You are participating in a decision-making study conducted by researcher Vytaute Rimkute. The study includes an effort task that you must complete within 90 seconds. If the timer runs out, you will be automatically redirected to the next page.

The task will ask you to move as many sliders to the requested position in a limited time (see Fig 1). There are 40 sliders overall. ***You can only move the sliders one after another. If any slider is missed, you will be disqualified****, and no earnings will be provided. If* ***any slider in a row, except the last one, is positioned incorrectly****, you will also be dismissed from the study.*

Since this is a tournament, the ***TOP 3 players will share a prize of $12*** *(If there are more than 3 players with a tie, then all these players, including the tie, will share the prize). After the study is completed and the performance of all participants is rated, the payments will be transferred to you via Prolific. There are a total of 150 participants competing in this tournament.*



12:29

Position the slider based on a **bolded number** above each slider.

Figure A.1: An Example of the Mobile Slider Task

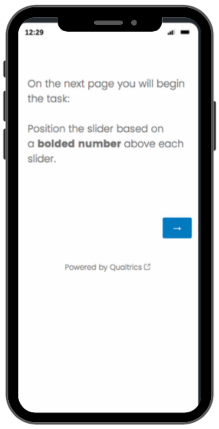
Your participation in the experiment is anonymous.

Good luck!

Appendix B Treatment conditions

Below, the different treatments are presented. The first screenshot refers to the baseline treatment, the second represents static treatment, and the last two screenshots depict the dynamic condition.

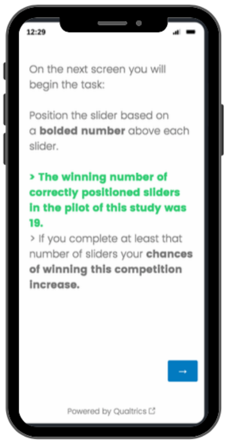
See Figure B1, Figure B2, and Figure B3.



On the next page you will begin the task:

Position the slider based on a **bolded number** above each slider.

Figure B.1: Baseline treatment



12:29

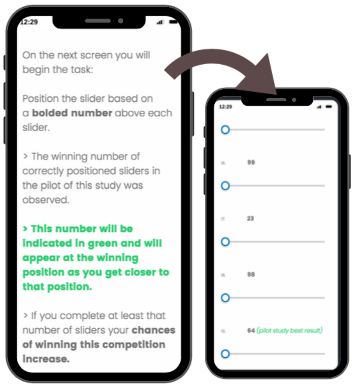
On the next screen you will begin the task:

Position the slider based on a **bolded number** above each slider.

> The winning number of correctly positioned sliders in the pilot of this study was 19.

> If you complete at least that number of sliders your **chances of winning this competition increase.**

Figure B.2: Static treatment



12:29

On the next screen you will begin the task:

Position the slider based on a **bolded number** above each slider.

> The winning number of correctly positioned sliders in the pilot of this study was observed.

> This number will be indicated in green and will appear at the winning position as you get closer to that position.

> If you complete at least that number of sliders your **chances of winning this competition increase.**

12:29

16. **99**

17. **23**

18. **98**

19. **64** *(pilot study best result)*

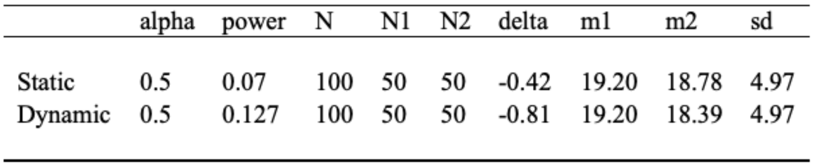
Figure B.3: Dynamic treatment

Appendix C Sample power estimates

The two sample power estimates for both devices and separately are provided below:

Table C.1: Total sample power for both devices based on correctly moved sliders

Estimated power for a two-sample means test t test assuming sd1 = sd2 = sd H0: m2 = m1 versus Ha: m2 != m1

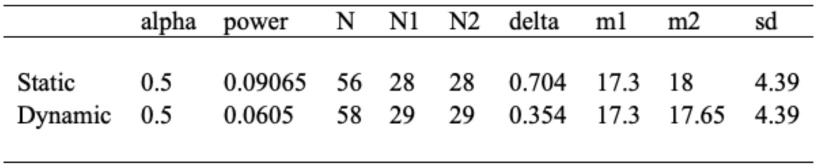


	alpha	power	N	N1	N2	delta	m1	m2	sd
Static	0.5	0.07	100	50	50	-0.42	19.20	18.78	4.97
Dynamic	0.5	0.127	100	50	50	-0.81	19.20	18.39	4.97

Note: N1 and N2 have been rounded down

Table C.2: Sample power for the iOS device based on moved sliders.

Estimated power for a two-sample means test t test assuming sd1 = sd2 = sd H0: m2 = m1 versus Ha: m2 != m1

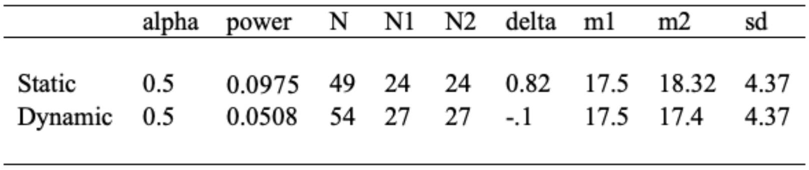


	alpha	power	N	N1	N2	delta	m1	m2	sd
Static	0.5	0.09065	56	28	28	0.704	17.3	18	4.39
Dynamic	0.5	0.0605	58	29	29	0.354	17.3	17.65	4.39

Note: N1 and N2 have been rounded down

Table C.3: Sample power for the iOS device based on correctly moved sliders (score).

Estimated power for a two-sample means test t test assuming sd1 = sd2 = sd H0: m2 = m1 versus Ha: m2 != m1

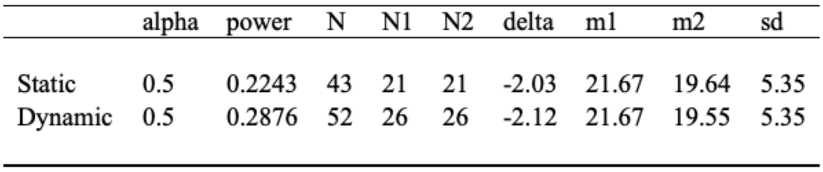


	alpha	power	N	N1	N2	delta	m1	m2	sd
Static	0.5	0.0975	49	24	24	0.82	17.5	18.32	4.37
Dynamic	0.5	0.0508	54	27	27	-.1	17.5	17.4	4.37

Note: N1 and N2 have been rounded down

Table C.4: Sample power for the Android device based on moved sliders.

Estimated power for a two-sample means test t test assuming sd1 = sd2 = sd H0: m2 = m1 versus Ha: m2 != m1

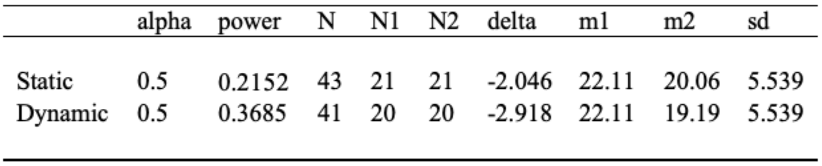


	alpha	power	N	N1	N2	delta	m1	m2	sd
Static	0.5	0.2243	43	21	21	-2.03	21.67	19.64	5.35
Dynamic	0.5	0.2876	52	26	26	-2.12	21.67	19.55	5.35

Note: N1 and N2 have been rounded down

Table C.5: Sample power for the Android devices based on correctly moved sliders (score).

Estimated power for a two-sample means test t test assuming sd1 = sd2 = sd H0: m2 = m1 versus Ha: m2 != m1



	alpha	power	N	N1	N2	delta	m1	m2	sd
Static	0.5	0.2152	43	21	21	-2.046	22.11	20.06	5.539
Dynamic	0.5	0.3685	41	20	20	-2.918	22.11	19.19	5.539

Note: N1 and N2 have been rounded down

Appendix D The OLS regression equation for the effect of treatment on performance.

Column 1 in the Table 3 estimates the following equation:

$$y_{it} = \alpha_0 + \beta_1 iOS_t + \varepsilon_{it} \tag{D.1}$$

where the dependent variable was the score of the total number of (correctly) moved sliders for individual *i*, with operating system *t*, and the estimate identifies the effect of the device (iOS) on performance.

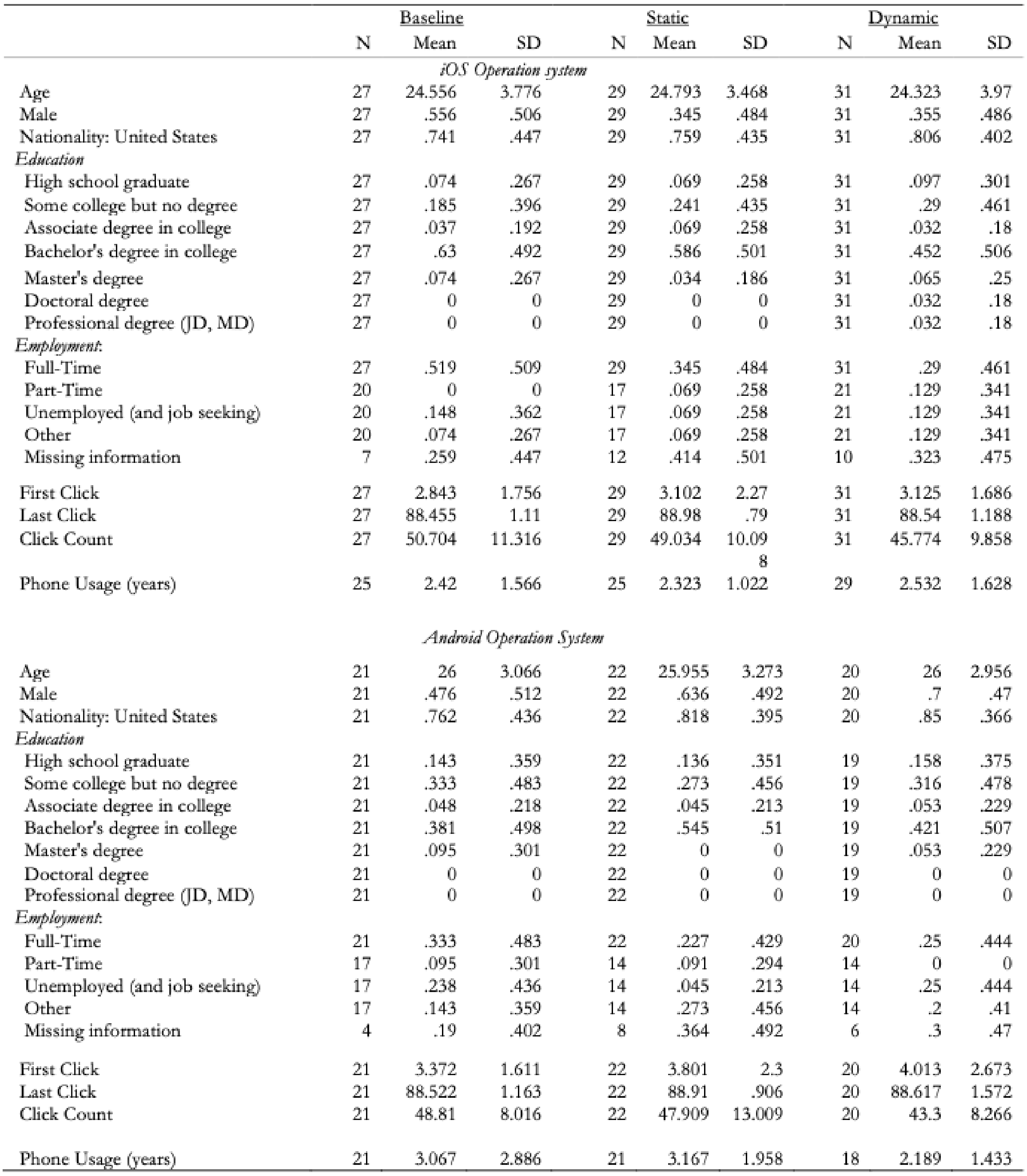
Column 2-8 in the Table 3 estimates the following equation:

$$y_{it} = \alpha_0 + \beta_1 Treatment_t + X_i\beta + \varepsilon_{it} \tag{D.2}$$

where the dependent variable was the score of the total number of (correctly) moved sliders for individual *i* in the treatment *j*, and β_1 estimate identifies the effect of the treatment condition. Where the vector controls for the device (Column 2-4) and browser (Column 5-7) effects.

Appendix E Summary statistics

Table E.1: Summary statistics by the device.



		Baseline			Static			Dynamic	
	N	Mean	SD	N	Mean	SD	N	Mean	SD
		iOS Operation system							
Age	27	24.556	3.776	29	24.793	3.468	31	24.323	3.97
Male	27	.556	.506	29	.345	.484	31	.355	.486
Nationality: United States	27	.741	.447	29	.759	.435	31	.806	.402
Education									
High school graduate	27	.074	.267	29	.069	.258	31	.097	.301
Some college but no degree	27	.185	.396	29	.241	.435	31	.29	.461
Associate degree in college	27	.037	.192	29	.069	.258	31	.032	.18
Bachelor's degree in college	27	.63	.492	29	.586	.501	31	.452	.506
Master's degree	27	.074	.267	29	.034	.186	31	.065	.25
Doctoral degree	27	0	0	29	0	0	31	.032	.18
Professional degree (JD, MD)	27	0	0	29	0	0	31	.032	.18
Employment:									
Full-Time	27	.519	.509	29	.345	.484	31	.29	.461
Part-Time	20	0	0	17	.069	.258	21	.129	.341
Unemployed (and job seeking)	20	.148	.362	17	.069	.258	21	.129	.341
Other	20	.074	.267	17	.069	.258	21	.129	.341
Missing information	7	.259	.447	12	.414	.501	10	.323	.475
First Click	27	2.843	1.756	29	3.102	2.27	31	3.125	1.686
Last Click	27	88.455	1.11	29	88.98	.79	31	88.54	1.188
Click Count	27	50.704	11.316	29	49.034	10.098	31	45.774	9.858
Phone Usage (years)	25	2.42	1.566	25	2.323	1.022	29	2.532	1.628
		Android Operation System							
Age	21	26	3.066	22	25.955	3.273	20	26	2.956
Male	21	.476	.512	22	.636	.492	20	.7	.47
Nationality: United States	21	.762	.436	22	.818	.395	20	.85	.366
Education									
High school graduate	21	.143	.359	22	.136	.351	19	.158	.375
Some college but no degree	21	.333	.483	22	.273	.456	19	.316	.478
Associate degree in college	21	.048	.218	22	.045	.213	19	.053	.229
Bachelor's degree in college	21	.381	.498	22	.545	.51	19	.421	.507
Master's degree	21	.095	.301	22	0	0	19	.053	.229
Doctoral degree	21	0	0	22	0	0	19	0	0
Professional degree (JD, MD)	21	0	0	22	0	0	19	0	0
Employment:									
Full-Time	21	.333	.483	22	.227	.429	20	.25	.444
Part-Time	17	.095	.301	14	.091	.294	14	0	0
Unemployed (and job seeking)	17	.238	.436	14	.045	.213	14	.25	.444
Other	17	.143	.359	14	.273	.456	14	.2	.41
Missing information	4	.19	.402	8	.364	.492	6	.3	.47
First Click	21	3.372	1.611	22	3.801	2.3	20	4.013	2.673
Last Click	21	88.522	1.163	22	88.91	.906	20	88.617	1.572
Click Count	21	48.81	8.016	22	47.909	13.009	20	43.3	8.266
Phone Usage (years)	21	3.067	2.886	21	3.167	1.958	18	2.189	1.433

Note: SD denotes standart deviation.

Appendix F Robustness checks

See Table F1, Table F2 and Table F3.

Extra calculations were required to determine whether individuals had seen the reference point.

The metadata includes the phone resolution and size (width and length). The size of one slider was approximately 162.4 pixels. Based on this data, I calculated the following:

1) The location of the reference points in terms of pixels:

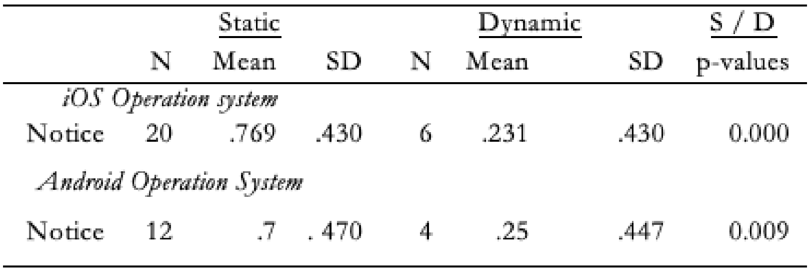
$$location = 19 \times 162.4 - length\,of\,the\,phone \tag{F.1}$$

2) The number of moved sliders had been obtained, and thus one could calculated what is the length of the total number moved sliders in terms of pixels:

$$total\,moved\,length = 162.4 \times number\,of\,total\,moved\,sliders \tag{F.2}$$

3) Lastly, a dummy could be constructed. If an individual's total number of moved sliders in terms of pixels is higher than the plausible location of seeing the reference point, one could assume that the reference point had been seen (coded as 1); if not, the dummy had been coded as 0.

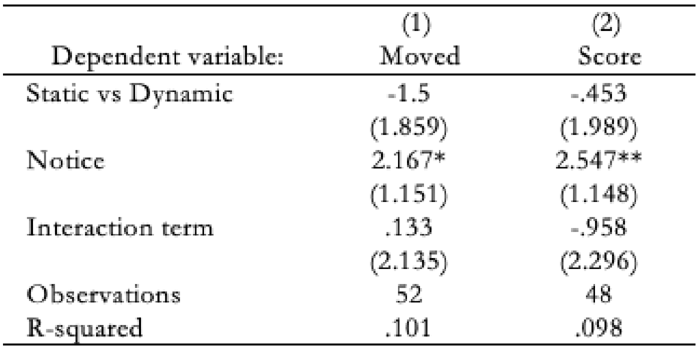
Table F.1: Non-parametric test results of the Fisher's Exact test for seeing the reference point.



	Static			Dynamic			S / D
	N	Mean	SD	N	Mean	SD	p-values
iOS Operation system							
Notice	20	.769	.430	6	.231	.430	0.000
Android Operation System							
Notice	12	.7	. 470	4	.25	.447	0.009

Note: SD denotes standart deviation.

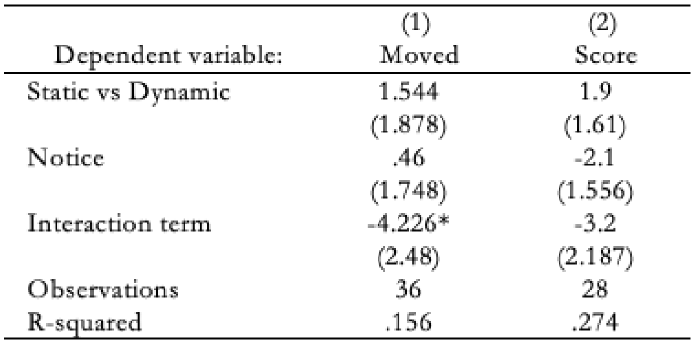
Table F.2: The effect of the reference point on performance for iOS devices and seeing the reference point.



Dependent variable:	(1) Moved	(2) Score
Static vs Dynamic	-1.5	-.453
	(1.859)	(1.989)
Notice	2.167*	2.547**
	(1.151)	(1.148)
Interaction term	.133	-.958
	(2.135)	(2.296)
Observations	52	48
R-squared	.101	.098

Note: Standard errors are in parentheses *** p<.01, ** p<.05, * p<.1

Table F.3: The effect of the reference point on performance for Android devices and seeing the reference point.



Dependent variable:	(1) Moved	(2) Score
Static vs Dynamic	1.544	1.9
	(1.878)	(1.61)
Notice	.46	-2.1
	(1.748)	(1.556)
Interaction term	-4.226*	-3.2
	(2.48)	(2.187)
Observations	36	28
R-squared	.156	.274

Note: Standard errors are in parentheses *** p<.01, ** p<.05, * p<.1

Appendix G Prolific

This was the description posted on the Prolific visible by respondents.

Slider game

ONLY MOBILE DEVICE PARTICIPANTS. In this study, you will take part in a tournament and will be playing a short game taking 90 sec. In addition to the flat rate fee, TOP3 players will share a price of $12. In case you have any question, please do not hesitate to send me an email vytaute.r@gmail.com .