Norwegian School of Economics
Bergen, Autumn 2015

Affecting the gender difference in risk-taking behavior

A study of how the gender gap in risk-taking behavior can be influenced by a default effect

Julianne Kallestad Øien and Siri Stenberg Østli

Supervisor: Alexander W. Cappelen

Master Thesis in International Business (INB) and

Business Analysis and Performance Management (BUS)

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

The purpose of this study is to investigate the possibility of influencing the gender difference in risk-taking behavior. By doing so, we combine research from two different fields, namely literature on gender differences in risk-taking behavior, and the literature on default effects. We wish to contribute to the research in the intersection between these two fields.

The question in focus is examined by gathering primary data through an incentivized economic experiment posted on the online platform Amazon Mechanical Turk. The 360 participants that contributed to our study were exposed to one of two treatment variations. Half the participants were initially given a default option encouraging risk seeking behavior, while the other half were given a default encouraging risk averse behavior. A randomized experiment enabled us to examine the causal relationship between the risk-taking behavior of the participants and the treatment they received.

In accordance with previous research and our expectations, our findings indicate greater risk aversion among females compared to males. Furthermore, we find evidence of a treatment effect, meaning that people receiving the risky option as default, are exhibiting more risk seeking behavior. Our most interesting result is found when interacting the treatment effect with gender. When dividing our sample by gender, we find a significant treatment effect among females, and no evidence of a treatment effect among males. This implies that only females seem to be affected by a default bias, while males are equally risk seeking irrespective of the default option.

In addition to the main analysis, we performed a short analysis of the influence of time preference on the tendency to stick with the default option, as we found this relationship interesting. Our findings imply that impatient individuals are significantly more influenced by the default effect than patient ones.

The result of our study substantiates the possibility of affecting the gender gap in risk-taking behavior by changing the default option of a choice. By framing the choice with the risky option as default, the gender gap disappears. This is interesting from a policy perspective, since it provides a different interpretation of the drivers of gender differences in risk-taking behavior. It suggests that the gender difference is not only an expression of underlying risk preferences. Rather, it might be a result of the combination of a stronger default bias among females, and that the safe alternative often is the default in many decisions that we meet in everyday life.

Preface

This paper is a master thesis written in the final year of our Master of Science in Economics and Business Administration at the Norwegian School of Economics (NHH). We specialize in the fields of International Business (INB) and Business Analysis and Performance Management (BUS). The thesis accounts for 30 credits within our majors. The topic of the thesis is within the field of behavioral economics, and the purpose is to investigate how risk-taking behavior can be influenced. Our aim is to examine whether the gender gap in risk-taking can be influenced by the way alternatives are framed, more specifically, by looking at the influence of a default effect.

The reason behind the choice of topic is our personal interest in behavioral economics. One of us was particularly interested in the gender differences in risk-taking and the other one in the biases in human decision making. Under the guidance of Professor Alexander Wright Cappelen, we wished to combine our interests and expand the knowledge about the possibility to influence risk taking. Few studies have looked at the implications of framing on gender differences in risk taking, and none have used the exact same framing as our study.

We express our gratitude to our supervisor, Professor Alexander Wright Cappelen, for valuable input and constructive feedback throughout the process. He was a great source of inspiration and motivation behind our research and topic. We believe that our interest and hard work along with Alexander Wright Cappelen's inspiration and enthusiasm have contributed to make this an interesting thesis. Hopefully, it provides a valuable contribution.

We also thank The Choice Lab for their financial contribution which made it possible to carry out the experiment. In addition, we thank Ingar K. Haaland and Ida Elisabeth H. Kjørholt from The Choice Lab for their good help with conducting the experiment. We have learned a lot during the process of writing this thesis, ranging from theoretical insights to methodological procedures for carrying out a scientific study.

Bergen, December 2015

Julianne Kallestad Øien

Siri Stenberg Østli

Content

List of tables

List of figures

1. Introduction and background

Risk fundamentally affects individual behavior and plays a crucial role in almost every important economic decision and numerous other non-economic decisions. Your risk preferences will affect several aspects of your daily life. It can influence everything from what kind of career you chose, to your propensity to drink alcohol or take drugs. Being more inclined to take risk is, in particular contexts, found to be associated with greater personal and corporate success (MacCrimmon and Wehrung, 1990).

It is well documented from economic experiments that most people are risk averse (Arrow, 1965; Pratt, 1964; Ross, 1981; Yates and Stone, 1992). Furthermore, research points towards females being more risk averse than males across different domains and contexts. These findings imply that there exists a gender gap in risk-taking behavior. The literature on default effects, including the more resent and increasingly popular theory on nudges, have proven that the influences of a default effect has considerable influences on people's choices and behavior. These results are consistent in settings ranging from insurance, investment and marketing to organ donations and health care (Johnson et al., 2012; Sunstein, 2014). The tendency of people to presume gains and losses relative to a reference point or the status quo can be seen as the mechanism behind the default effect.

In this master thesis we take a closer look at the default effect, how it works and who it influences, and combine this with the findings from the research on risk-taking behavior. We have built upon existing research to formulate a hypothesis about what we expect to find. The effects of a default on the gender difference in risk-taking behavior were tested by designing and executing an online economic experiment which provided primary data. Primary data gives control over both the data obtained from the respondents and the sample structure. This increases confidence that the data will match the objectives of the study (Easterby-Smith, Thorpe and Jackson, 2008).

To be able to measure risk-taking in the most realistic and reliable way, we made use of an incentivized experiment with real money at stake. This is the prevalent way of measuring risk-taking behavior in today's research, and strengthens the validity of our results. The sample, consisting of 360 participants, where randomly assigned to one of two treatment variations.

Thereafter, they were given the task intended to reveal their risk-taking behavior. This task was a choice between a risky payment option (a lottery) and a safe payment option (a certain/ fixed amount of money). One treatment group was given the risky option as default, whereas the other group was given the safe option as default.

Our main finding is that females and males respond differently to the treatment. In the treatment group with the safe default, a greater share of women than men chose the safe option. This is consistent with previous findings claiming that women exhibit more risk averse behavior. However, in the treatment group with the risky default, the gender differences in risk-taking is neutralized, meaning that the same amount of women and men choose the risky option. Men seem to be unaffected by the default effects as there were no differences in risk-taking across treatments.

The findings indicate that female risk-taking is affected by the default option, whereas the risk-taking of males is unaffected by the same effect. Hence, it might seem like the gender differences in risk-taking behavior is more a question of loss aversion and reference dependence (default effect), than a question of underlying preferences for risk. It is possible that women are not more risk averse than men, but possess a stronger default bias. This is a plausible interpretation if most choices, in general, are framed with a safe default rather than with a risky default. Consequently, our study emphasizes the importance of considering the framing of choices.

Several studies have looked into differences in risk-taking when it comes to the framing of choices. However, the research on gender differences in default bias is very limited. Our study contributes to this literature and to the understanding of gender differences in risk-taking behavior by documenting a strong relationship between the default effect and female's propensity to take risk.

Furthermore, we performed a brief additional analysis to investigate the relationship between the default effect and people's time preferences. During our analysis we found this variable particularly interesting as the treatment effect seemed to be partly explained by time preference rather than gender. Further analysis revealed two main findings related to time preference. Firstly, we found impatient individuals to be much more risk averse than patient ones, significant at the 99 % level. Secondly, we found impatient individuals to be more affected by the default option than patient ones, significant at the 90 % level.

1.1 Background and motivation

The topic of this thesis started as an interest in why there are less female leaders than male leaders. There exist several hypotheses seeking to explain this phenomenon. One being that women are still not given the same opportunities as men, rooted in a history of gender inequality. Another states that women are kept out of top management because they will be absent in periods when they are having children; thus making men appear a safer choice due to continuous availability. It has also been argued that women are less inclined to make career sacrifices origin from a higher sensitivity to work-family conflicts (Gneezy, Leonard and List, 2009).

Several studies in the experimental economics literature have suggested that men are more competitively inclined than women (e.g. Almås, Cappelen, Salvanes, Sørensen and Tungodden, 2012; Gneezy et al., 2009; Niederle and Vesterlund, 2007; Schurchkov, 2012). Another explanation put forward by Schurchkov (2012) is that gender differences in skills and preferences lead to occupational self-selection. Yet another one is that women are choosing other careers because a career in top management involves significant risk. Such a career comes with a large responsibility and consequently a large risk of making bad decisions and mistakes. The potential fall is much larger from the "top of the ladder".

The hypothesis that women are more risk averse than men has been put forward as a major cause behind the "glass ceiling" (Johnson and Powell, 1994). The "glass ceiling" can be described as "the unseen, yet unbreakable barrier that keeps minorities and women from rising to the upper rungs of the corporate ladder, regardless of their qualifications or achievements" (Federal Glass Ceiling Commission, 1995, p. 4). Some argue that women will not make the risky decisions that might be necessary for a business to succeed (Schubert, Brown, Gysler and Branchinger, 1999). Eckel and Grossman (2008, p. 2) provides a description of the importance of risk preferences:

> *Whether men and women systematically differ in their responses to risk is an important economic question. If women are more sensitive to risk than men, this will be reflected in all aspects of their decision making, including choice of profession (and so earnings), investment decisions, and what products to buy.*

In addition to the potential relationship between risk-taking behavior and female under-representation in top management positions, risk-taking is important in several other aspects of business. Risk preferences can be decisive when firms are hiring, or in other situations where people might be selected based on their risk preferences. For instance, startup companies may be looking for risk seeking employees when expanding, and investment managers might need the right risk preferences to be assigned to important clients (Weber, Blais and Betz, 2002). There has also been a debate whether risk averseness cause fewer women to become entrepreneurs. Several studies find that women are less likely to engage in entrepreneurial activities than men (e.g. Zeffane, 2013).

1.2 Research question and structure of the thesis

We wish to look into the gender differences in risk-taking behavior. At the same time, we want to investigate whether risk-taking behavior could be influenced by a small change in the formulation of a question. The focus of our study will be on the simultaneous influence of these two effects. The aim is to reveal the potential relationship between gender differences in risk-taking and gender differences in default effects. We define our research question in the following way:

Research question: *What are the effects of a default option on the difference in risk-taking behavior among men and women?*

The thesis is structured in five main chapters. The first chapter is the introduction where we explain the motivation and background for the study. The second chapter provides a literature review, where we examine the existing literature in the relevant fields, to understand what is known and what is not known about our chosen topic. At the end of chapter two we develop a hypothesis to our research question based on existing theories and previous research. Chapter three presents the method used for investigating our research question. This chapter offers a description of how our research was conducted and justifies the methods used and the choices made. Chapter four presents the findings and results of our study. We made use of statistical tests to analyze the data in order to answer our research question. In the fifth and last chapter, a brief summary of our study is offered and we discuss our findings and their implications. At the end of this chapter, we present some suggestions for future research, and conclude our study.

2. Literature review

In this section, we present a review of the relevant theories, concepts and empirical studies in the current literature on behavioral economics. This will provide an overview and understanding of the most important knowledge and also the latest findings in the field. Existing research will help us formulate a hypothesis to our research question, which is proposed at the end of the literature review. The review is structured in three main parts. First, we present the relevant literature and findings to understand risk-taking behavior and gender differences in risk taking. Thereafter, we describe the concepts of reference dependent preferences and loss aversion, within the prospect theory framework. Finally, we present our hypothesis.

2.1 Risk-taking behavior and gender differences

The study of risk-taking behavior is a multidisciplinary exercise. Definitions of *risk* and *risk-taking* depends on the field of study and varies between economics, management sciences, psychology, anthropology and sociology (Shapira, 1995). The dominant theory of decisions under risk is the expected utility theory. In this framework, risk attitude is a feature of the shape of the utility function that underlies a person's choices (Weber et al., 2002). A utility function can be graphed in a diagram with utility on the y-axis and something of value on the x-axis, such as wealth, income or money, see Figure 1 below.

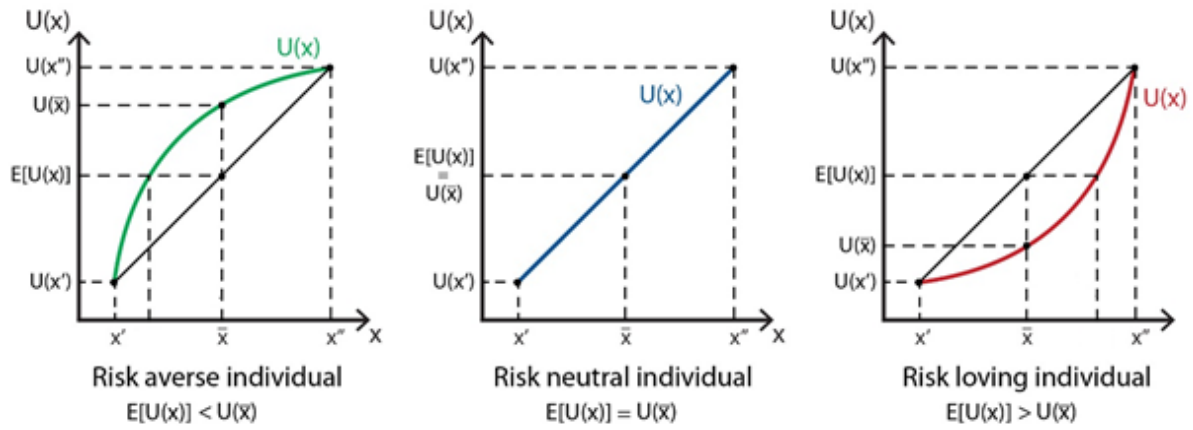


Figure 1: Risk preferences (Source: Policonomics, 2012)

People who are less willing to take risk are often described as *risk averse* and have a concave utility function (left-hand graph). They will gain less utility from an uncertain option with an expected value of x, than from an option with a certain value of x. This can be illustrated as a choice between an uncertain option where one could win $10 or nothing with equal probabilities and a certain payment of $5. The expected value of the uncertain option is $5 ($10*0.5 + $0*0.5). A person who is indifferent between this certain and uncertain option, is *risk neutral* (middle graph). For these people the utility of an uncertain option with expected value x, is equal to the utility of a certain payment of x.

People who are more willing to take risk are often described as *risk seeking* and have a convex utility function (right-hand graph). They could gain utility from accepting an uncertain option even if the expected value is below the certain payment (Mongin, 1997). Yates and Stone (1992) states that the "pure" attitude towards risk is always negative and people will require a premium in return to take on risk. This relationship was established several decades ago in studies by Pratt (1964), Arrow (1965), Ross (1981) and others who all find individual human decision makers to be risk averse.

There is a large literature on gender differences in risk-taking behavior. Not only within the field of economics, but also within sociology and psychology. The studies are conducted in numerous ways and contexts and provides a solid foundation for discussion. We will focus on the studies framed in an economic context.

One of the largest meta-analysis on this topic was conducted in 1999 by Byrnes, Miller and Schafer. They compared 150 separate studies and classified 16 different types of risk taking. In 14 of the 16 different types, they found that there were a significant difference between males and females, with females being more risk averse. The differences where larger in certain contexts, such as intellectual risk taking, and smaller in others. The study also focused on gender differences as a variable of age, and found that the gender gap did vary with age, although the direction and magnitude depended on the context.

More recent meta-analyses also find the same results. Charness and Gneezy (2011) gathered data from 15 sets of experiments that were not designed to investigate gender differences. All experiments are based on the same underlying investment game, but they were conducted by

different researchers in different countries and with different subject pools. This strengthens the conclusion that females are more risk averse than males.

Croson and Gneezy (2009), compare several studies using both real and hypothetical gambles. Their robust findings are that men are more risk prone than women in both lab settings and investment decisions in the field (real life setting). One example of such a field study is Sunden and Surette's (1998) investigation of asset allocation. They find that women invest their assets more conservatively than men do. Bajtelsmit and VanDerhei (1997) also find that a large percentage of women invest in the minimum-risk portfolio available to them.

Eckel and Grossman's meta-study from 2008 found some differing results, depending on the type and framing of the independent studies. They separate experimental studies, where the subjects know that they participate in a study, from field studies, where subjects are observed in real life settings. Within the experimental studies, abstract experiments are separated from contextual environments. With an abstract environment, participants could for instance be asked to choose between a lottery and a safe payment. With a contextual environment, you add context to the experiment, framing the choices differently. In this case the participants might be asked to allocate their investment between a safe and a risky asset.

Lastly, Eckel and Grossman distinguish between studies that conduct experiments in the gain domain and in the loss domain, referring to the framing of the games. As an example of an experiment in the gain domain one could ask the participants to choose between a safe payment of 10 dollars or a lottery where you can win 30 dollars or noting with equal probabilities. If this game were to be in the loss domain one could ask the participants to choose between losing 10 dollars with certainty or a lottery where you lose either 30 dollars or nothing with equal probabilities.

The first study investigated in Eckel and Grossman is Brinig (1995). Brinig conducts abstract experiments in the gain domain, and does not perform experiments in the loss domain or in contextual environments. Brining finds that when gender is interacted with age, it becomes a significant predictor of risk taking. The difference in risk-taking peaks at about age 30. This could be explained by the tendency of men to be more risk seeking during the period when they are trying to attract mates. Women on the other hand are more risk averse during the

period when they usually have children. Both these periods often occur in the years around age 30. Several other studies also confirm that age has an impact on willingness to take risks.

Most studies in the meta-analysis by Eckel and Grossman find females to be significantly more risk averse than males. Some studies are inconclusive and cannot find a significant difference between the genders. Only two studies find men to be more risk averse, and both of them are abstract experiments in the loss domain. Schubert et al. (1999) presents their subjects with four choices between certain payoffs and risky lotteries. Two of the choices are framed in an abstract environment, and the two others in a contextual environment. Within each environment, one choice is presented as a potential gain and the other as a potential loss. In the contextual environment, the results are inconclusive. In the abstract environment, the results are significant. However, the results are reversed from the gain domain to the loss domain, stating that females are most risk averse in the gain domain, while men are more risk averse in the loss domain. These are interesting results, although not unambiguously supported by other research.

In summary, Eckel and Grossman are rather conservative in their view, stating that "the findings thus far shed serious doubt on the existence of risk attitude as a measurable, stable personality trait". They argue that it is difficult to make any conclusion about the gender difference. This stands in contrast to Charness and Gneezy (2012), who named their comparative analysis *Strong Evidence for Gender Differences in Risk Taking.* Evidently, the researchers are not unanimous in regards to the evidence of gender differences. Literature tends to either find females to be more risk averse, or no significant gender differences. Studies finding females to be more risk averse are robust across different contexts, especially in the gain domain.

2.2 Reference dependent preferences and loss aversion

Prospect theory is a framework that formalizes the idea of loss aversion and reference dependence preferences. This theory was presented as a critic against the expected utility theory, which was the dominated model of decision making under risk (Kahneman and Tversky, 1979). Kahneman and Tversky (1979) base their critic on empirical evidence revealing that people are not always rational or consistent when it comes to their preferences. In particular, people have a tendency to overweight outcomes that are obtained with certainty

to outcomes that are merely probable. This so called *certainty effect* contributes to risk seeking behavior in choices involving sure losses and risk aversion in choices involving sure gains. In addition, people tend to simplify choices between alternatives by focusing on components that differentiates them. Thus, people often disregard components that are shared among the alternatives. This *isolation effect* may produce inconsistency in people's preferences when the same choice is presented in different ways (Kahneman and Tversky, 1979).

Compared to expected utility theory, prospect theory assigns value to gains and losses rather than to final assets, and replaces probabilities with decision weights. The prospect theory model is outlined in Figure 2 below. Three essential features characterize the value function: *reference dependence, loss aversion and diminishing sensitivity.* The location of the reference point has implications for the perception of an outcome as a gain or a loss. Loss aversion is the tendency of people to prefer avoiding losses to acquiring gains. A loss will have a greater negative impact on satisfaction, than the positive impact of a gain of the exact same amount (Tversky and Kahneman, 1991). The concept of diminishing sensitivity is that "marginal value of both gains and losses decreases with their size" (Tversky and Kahneman, 1991, p. 1039). These three properties creates an S-shaped value function that is convex below the reference point and concave above it.

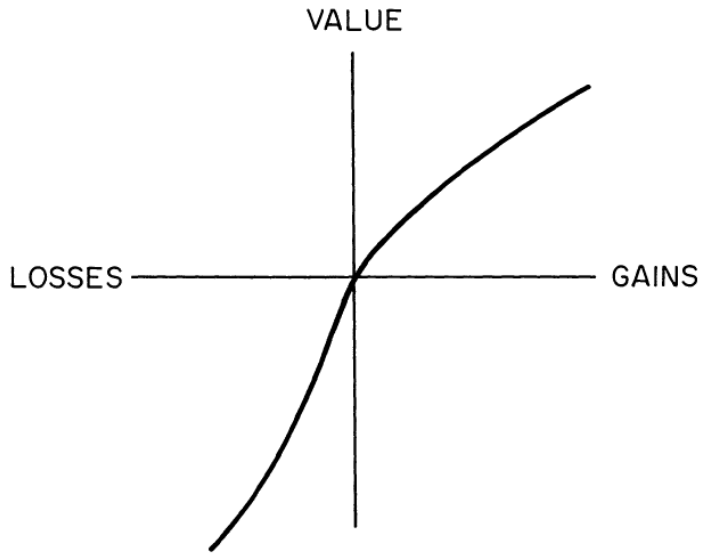


Figure 2: An illustration of a value function (Tversky and Kahneman, 1991, p. 1040).

The concavity of the value function entails risk aversion in the gain domain, whereas the convexity entails risk seeking in the loss domain. Risk seeking in the loss domain has by several investigators been confirmed and seems to be accepted in the literature as rather robust (e.g. Fishburn and Kochenberger, 1979; Hershey and Schoemaker, 1980; Payne, Laughhunn, and Crum, 1980; Slovic, Fischhoff, and Lichtenstein, 1982). This has further been observed by studies using nonmonetary outcomes (e.g. Fischhoff, 1983; Tversky and Kahneman, 1981). Evidence suggests that being risk seeking in the loss domain is especially strong when the probabilities of loss are substantial (Kahneman and Tversky, 1984).

However, some studies have questioned prospect theory's applicability. One study conducted by List (2004), which look into prospect theory vs. neoclassical theory in the marketplace, states that prospect theory falls short of predicting people's behavior when they are experienced. They point to the fact that people, when experienced, approaches the neoclassical prediction. This implies that prospect theory may only be applicable to inexperienced people or consumers. However, it is important to notice that the prospect theory only helps economist to explain what people do in certain situations, it does not substitute the idea of revealed preferences.

The idea behind prospect theory is that people are loss averse and that they experience gains and losses relative to some reference point. This can explain several important social phenomena (Levin, Schneider and Gaeth, 1998). More specifically it can explain the concepts of framing, status quo bias, the endowment effect and last but not least the default effect.

Prospect theory proposes that *framing* or phrasing matters. The way alternatives are framed or presented has by empirical testing been demonstrated to have big influences on people's choices (e.g. Tversky and Kahnmeman, 1991). This is illustrated in a classical experiment, named "the Asian disease", by Tversky and Kahneman in 1981.

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume the exact scientific estimate of the consequences of the programs as follow:

The first treatment group was asked to choose between program A and B:

1. *If program A is adopted 200 people will be saved (72%)*
2. *If program B is adopted, there is one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved (28%)*

The other treatment group was asked to choose between program C and D:

1. *If program C is adopted, 400 people will die (22%)*
2. *If program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die (78%)*

In this experiment the reference point is that everyone will die, which is stated in the introduction. Results shows that when the options are framed positively 72% of the sample chose the risk averse option (alternative A), whereas only 22% chose the equivalent option (alternative C) when the options are framed negatively. Hence, people seem to choose alternatives that are framed positive rather than negative, depending on the reference point. In other words, people tend to experience loss aversion.

The status quo concept is related to the reference point. One implication of loss aversion is the tendency of people to stick with the current situation (Kahneman, Knetsch and Thaler, 1990). The mechanism behind loss aversion causes us to be more afraid of the potential losses of switching from status quo than the potential gains. However, in the absence of loss aversion several other factors can induce a status quo bias (Tversky and Kahneman, 1991). Samuelson and Zeckhauser (1988) offers several explanations to this bias, including inertia, cost of thinking, fear of regret in making a wrong decision, transaction costs, perceiving what have worked in the past as a safe option, and avoiding the need to make an active choice, i.e. preference of doing nothing.

A series of decision-making experiments show that individuals often stick with the status quo alternative (e.g. Samuelson and Zeckhauser, 1988). In their experiments, Samuelson and Zeckhauser (1988), frame a hypothetical situation by giving one group an alternative as a status quo, rather than provide all the alternatives as options, as opposed to the other group. Their results imply that, the framing of an alternative, whether it is in the status quo position or not, significantly affect the likelihood that the alternative will be chosen. The robustness of status quo bias is enhanced by field studies revealing consistent results (e.g. Hartman, Doane

and Woo, 1991; Madrian and Shea, 2000). This implies that status quo bias is important in "real world" decisions and not only within an artificial laboratory setting.

The endowment effect, which is closely related to the concept of status quo, is a bias that make people stick with the current option. Loss aversion affects people to value a good more when the good becomes part of their endowment, i.e. when people own the good themselves (Kahneman et al., 1990). Several researchers have demonstrated the prevalence of the endowment effect (e.g. Knetsch and Sinden, 1984, Knetsch, 1989; Loewenstein and Kahneman, 1991). In Kahneman, Knetsch and Thaler's (1990) study, half the participants were given a coffee mug and the opportunity to sell it. The other half were given the opportunity to buy a coffee mug. With no endowment effect, the price people would be willing to accept for selling the mug should be similar to the price people are willing to pay for the mug. However, this is not the case. The "willingness to accept" was about twice as high as the "willingness to pay". Once people had established an ownership of the mug, they valued it much higher. In spite of this, List (2004) argues that "the consumer learns to overcome the endowment effect in situations beyond specific problems they have previously encountered" (List, 2004, p. 615). This is consistent with List's criticism toward prospect theory's applicability to experienced consumers.

Brown and Krishna (2004, p. 529) defines a default option as "the one the consumer will automatically receive if he/she does not explicitly specify otherwise". Hence, a default can be described as the tendency to stick with the status quo (reference point). However, there are other mechanisms giving power to the default option. When someone sets a default, many people might believe, wrongly or rightly, that the default is set that way it is for a reason. They might perceive the default as an implicit endorsement or recommendation from those who chose the default (Sunstein and Thaler, 2008). One last mechanism worth mentioning is laziness and procrastination, which in several contexts can be a main reason for the effectiveness of a default setting (Cronqvist and Haler, 2004). When people do not bother to make a choice or always intend to make the choice tomorrow, the default setting will be chosen for them.

Default options have proven to be effective and powerful in many settings ranging from insurance, investment and marketing, to organ donations and health care (Johnson et al., 2012). For instance, if people are automatically enrolled in retirement plans, their savings can increase significantly (Sunstein, 2014).

A default choice is listed as one of the most important and efficient types of "nudges" (Sunstein, 2014). A Nudge is an idea born in the US and popularized by Richard H. Thaler and Cass R. Sunstein through their best-selling book "Nudge" published in 2008. It is based on the idea that small and apparently insignificant details can have major impacts on people's behavior. In their book, Sunstein and Thaler (2008, p. 6) define *nudge* in the following way:

> *A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.*

Choice architecture can be understood as the context in which people make decisions. This could be the way the choice is presented or the order of the alternatives. Almost everything that could affect people's choices can be called a nudge. One example is the placement of healthy food in grocery stores and canteens. If it is placed where it is easy to see and pick, it is more likely that people would choose it, as opposed to if it was placed somewhere more hidden. However, it would not count as a nudge if it constrained people's alternatives. Banning junk food cannot be called a nudge (Sunstein and Thaler, 2008).

Other examples of nudges include reducing the size of the plate to make people eat smaller portions, reminding people about the health consequences of smoking, displaying the number of calories in meals, automatic enrollment in a pension plan, double-paged printing as a default setting and a text message to remind people about their doctor's appointment for the next day (Sunstein and Thaler, 2008).

One study, conducted by Heijden, Klein, Müller and Potters (2011), provides some interesting findings about personality and the influence of a nudge. They were interested in people's time preferences (or discounting rates) and the relationship between time preferences and nudging. They conducted a study using a sample of 1102 Dutch individuals were they nudged individuals to evaluate risk in combination instead of in isolation. The result showed that impatient individuals are more "nudgeable" than patient ones. These results are important because, as Heijden et al. describes, "impatient individuals are often the target group of nudges as impatience is associated with problematic behaviors such as low savings, little equity holdings, low investment in human capital, and an unhealthy lifestyle" (Heijden et al.,

2011, p. 1). They concludes their study by encouraging more research on the topic, and especially on the effects of other types of nudges. They express that "it would be interesting to see whether default effects or the impact of commitment devices are stronger for impatient than for patient individuals." (Heijden et al., 2011, p. 17).

There is not much available literature on gender differences in risk taking when it comes to the direct influence of a default effect. We could only find one study touching upon this question. Agnew, Anderson, Gerlach and Szykman (2008) investigated the gender difference in framing effects, including default effects. Their study differs from ours in several aspects, for instance that are they using a more contextual environment. Nevertheless, their results could provide some indications of what to expect. They studied framing in the context of retirement savings, and gave their subjects the option between purchasing an annuity (safe option) and investing their savings on their own (risky option). They look into two different framing effects, namely attribute framing (or negative framing), and default effects.

The attribute framing was given as a five minute slide-show highlighting the negative aspects of one of the options. They find women to be influenced by the negative framing of investments, whereas men are influenced by the negative framing of both options. Regarding the default, participants were either given the investment, the annuity or neither of the two as a default option. They find that giving the investment as default has no significant effect on neither of the genders. Giving the annuity as default influenced male risk taking in one of their regression models, but not in the two other models. Female risk taking was not influenced by any default. These findings stand in contrast to the prevalent results in the literature, which find defaults to influence decision making. The authors argue that this might be the case because their default is too weak to cause an effect. We could not find any other studies investigating the gender differences in default bias. However, more studies have look at gender differences when it comes to other types of framing effects.

Another study conducted by Hasseldine and Hite (2003) use *goal framing* by manipulating "two objectively equivalent messages (one positively framed, one negatively framed) that are communicated to adult taxpayers" (p. 517). They find that women are more influenced or persuaded when messages are positively framed, whereas men are more persuaded when negatively framed. Regarding goal framing, several other studies find that negatively framed messages are more persuasive than positive ones independent of gender (e.g. Ganzach and

Karsahi, 1995; Meyerowitz and Chaiken, 1987). However, one should note that Hasseldine and Hite (2003) find females to be less interested and experienced with tax matters.

A third framing effect is characterized as *risky choice framing*, and an example of this is the "Asian disease" experiment outlined above. Fagley and Miller (1990) look into experiments that have revealed different results in the "Asian disease" and find that women are influenced by framing. Men, on the other hand, seem to be consistent in their preferences independent of framing. These results across tasks domains (attribute, goal and risky choice framing) indicate that gender differences in framing effects exist, although highly context specific. The framing effect seems to be influenced by interest and experience (Hasseldine and Hite, 2003). We therefore highlight the necessity to separate individual decision making in different task domains when investigating framing effects (Huang and Wang, 2010).

Regarding gender differences in status quo bias, endowment effect or loss aversion, we find one study that looks into possible gender differences in loss aversion. Schmidt and Traub (2001) conducted an experimental test of loss aversion. They find significant results implying that women experience loss aversion more frequently and to a higher degree than men. This finding implies that loss aversion may only be an important factor for some people, in this case females. They report that female subjects substantially contribute to their finding that people experience loss aversion in decision making. "Thus, the conclusion may be drawn, that women have a higher degree of risk aversion than men at least partly because they are more loss averse" (Schmidt and Traub, 2001, p. 18).

Considering that the literature did not have much to offer on which gender is most affected by the default option, we turned to other related literatures to try to find research that could help us answer our research question. Although it might not be directly related, it could help us formulate our hypothesis. A meta-study by Eagly and Carli (1981) find that women are more easily influenced in general than men. They also find women to be more persuadable and conforming than men. As this meta-study includes 148 separate studies, making the results more reliable, this might indicate that women are more influenced by a default effect than men.

2.3 Combining risk-taking behavior and the default effect

We have now reviewed the relevant concepts and empirical findings in the current literature on risk-taking behavior and the default effect. Our study will combine these two literatures by investigating the relationship between the gender difference in risk-taking and the influence of a default effect. We will look into whether the gender difference might be caused by the framing of choices. If most choices are framed with a safe option as default, this might partly explain the observed gender gap.

Most studies on risk-taking have found females to be more risk averse than males. We presume that we will find similar results. Regarding the default effect, studies have shown that people's choices are affected by a default option. However, there is limited research on gender differences. The only study found discovered no gender differences in default bias (Agnew et al., 2008). Schmidt and Traub (2001) find women to be more biased by loss aversion, and Fagley and Miller (1990) find that women are influenced by framing although men does not seem to be influenced. Eagly and Carli (1981) find women to be more easily influenced than males in general. Based on these indications in the literature, we find it reasonable to believe that females will be influenced by the default to a larger degree than males. This means that when exposed to a safe default, the gender gap in risk-taking will increase. When exposed to a risky default, the gender gap in risk-taking will be narrowed. To summarize our expectations, we define our hypothesis as follows:

Hypothesis: *The gender gap in risk-taking is larger with a safe default than with a risky default.*

The following illustration demonstrates our research hypothesis:

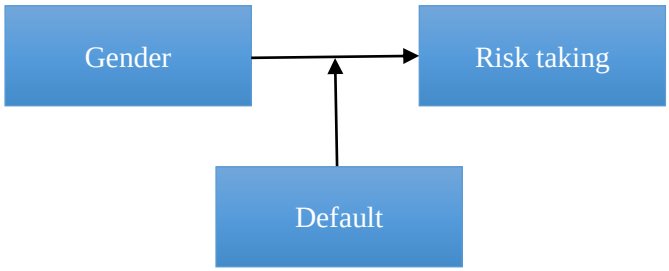


Figure 2: Research hypothesis

Based on previous research we assume that gender has an influence on risk taking. Females exhibit more risk averse behavior and males exhibit more risk seeking behavior. Furthermore, we expect that a default, or reference point, will moderate the effect of gender on risk taking. If the default is set to a safe option, the gender effect will be amplified, making the difference between males and females larger. If the default is set to a risky option, the gender effect will be dampened, making the gender difference smaller.

3. Methodology

The purpose of our study is to investigate the influence of a default effect on the gender gap in risk-taking behavior. We intend to explain the causal relationship between these variables using statistical analyses. To test and focus evidence about causal relationships, an experiment will be the most appropriate research strategy. Within this choice lies an implicit assumption that reality can be measured by numbers and analyzed with statistical techniques (Jacobsen, 2000). An experiment is ideal for a casual effect study because it helps eliminating other factors influencing risk-taking by comparing similar groups, where the only difference is the treatment itself or chance[1] (Haslam and McGarty, 2004).

In this chapter we present our experimental design. The chapter is divided in two subchapters. The first subchapter explains the design and structure of the experiment by describing the three main parts of the experiment in chronological order. The next subchapter explains how the experiment was carried out. This section also offers a description of the platform used to recruit the participants and a presentation of our obtained sample.

3.1 Design of the experiment

The experiment consists of three main parts: the work task, the part where we measure risk-taking behavior and the background questions. In addition, there is an introduction where the participants are informed about confidentiality, duration, payment and that participation is completely voluntary. In the next three sections we describe each of the three main parts in detail and justify why the different parts are included in the experiment. The fourth and last section of this subchapter provides an overview of the experiment and the effects we want to investigate. The complete survey is presented in Appendix A.1.

3.1.1 Part one: Work task

The first component of the experiment is a work task that all participants have to complete. This part is included to make the participant feel more entitled to the bonus they receive later on. By making people "work" for their bonus, we seek to increase the feeling that they

[1] "By chance "significant" findings may occur. For example, by *chance* 5 out of 100 correlation coefficients are expected to be significant at a 5 % level" (Ghauri and Grønhaug, 2010)

deserve the bonus, thereby making it appear more valuable. The specific task is a picture categorization task, where the participants are asked to choose the elements that best describes a given picture. They have to work on each picture for 30 seconds before they are automatically given a new picture. There are 6 different pictures, implying that the work task will last for 3 minutes in total.

3.1.2 Part two: Measuring risk-taking behavior

The second part is the main part of the study. This is where we measure the participants' risk-taking behavior. Risk-taking behavior is the dependent variable in our study and needs to be operationalized into a factor that can be empirically measured. We have chosen a simple and clean-cut design where the participants are given a choice between a safe payment and a lottery ticket. In this context, the action of choosing lottery indicates a risk seeking behavior and a higher preference for risk. Choosing the safe payment indicates a risk averse behavior and a lower preference for risk. This experimental design is chosen due to its predicative power of risk-taking behavior and its prevalence as a measurement in existing research (e.g. Harbaugh, Krause and Vesterlund, 2002; Schubert, et al., 1999).

In specific, the choice is structured as follows: the participants get the option between a safe payment of 1 USD and a lottery with the possibility to win 2.5 USD or nothing, with equal probabilities. The two alternatives, safe payment and lottery, would have the same expected value if the safe payment was 1 USD and the lottery was 2 USD or nothing with equal probabilities[2]. If people were risk neutral, they would be indifferent in the choice between these two alternatives. However, as most people are risk averse rather than risk neutral, they would gain more utility from choosing the safe payment. Ideally, we would want to offer alternatives that would make equally many participants chose the safe payment and the lottery. Hence, we have to make the potential gain in the lottery larger than 2 USD. A review of previous studies on risk-taking behavior showed that a multiple of 2.5 appeared to be a reasonable ratio (e.g. Charness and Gneezy, 2001).

We wanted to measure risk-taking through an incentivized choice with real money at stake because this is believed to yield more reliable and valid results than other measures. It is likely to assume that the probability of people answering in line with their true preferences is

[2] Expected value of lottery: 2 USD * 0.5 (probability to win) + 0 USD * 0.5 (probability to lose) = 1 USD

increased when using real money, rather than a hypothetical lottery. Measuring risk-taking by asking the respondents directly about their behavior would not be an ideal method. Economists are skeptical about whether self-reported attitudes will reflect actual risk-taking behavior. Various factors, including self-serving biases, inattention, and strategic motives could cause respondents to distort their reported risk attitudes and behavior (Dohmen et al, 2011). As a consequence, researchers rather rely on experimental measures of risk-taking behavior with real money at stake.

3.1.2.1 Treatment variation

In our experiment, participants are randomly assigned to one of two treatment groups, who each receive one of two different stimuli. The two groups will be similar with regards to all relevant aspects of the research, except the manipulation they receive. One treatment group will be given the lottery ticket as a default payment option, and thus exposed to a stimuli encouraging risk seeking behavior. The other treatment group will be given the safe payment as a default payment option, and thus exposed to a stimuli encouraging risk averse behavior. Regardless of their initial endowment (default), all participants get the opportunity to keep or exchange their received payment option.

According to standard economic theory, assuming that humans are rational decision makers, a default option will not influence behavior. In this framework, people make consistent choices in line with their underlying preferences regardless of the framing of options. As opposed to this theory, we want to create a reference point or status quo for the participants through the default option provided. In the two treatment groups we intend to exploit people's tendency to evaluate gains and losses relative to the status quo (reference dependence). In addition, we utilize their tendency to ascribe more value to things they own (the endowment effect) and their propensity to perceive a loss of something they own as much more powerful than gaining something they do not own (loss aversion). As explained in the literature review, these are the main effects (or biases) that make people stick to their initial endowment, rather than exchanging it.

In some cases, the effectiveness of a default is caused by the fact that people are not aware that they have a choice, or because the transaction cost of changing from the default are too high. In our study, we do not want this to be the reason for the effectiveness of the default.

Consequently, we intend to make it clear to the participants that they have a free choice. We make it easy for them to choose the alternative they want by eliminating transaction costs. The default option is implemented by slightly changing in the wording of the question provided to the two groups. The potential problem with this design is that the default might be too weak to cause an effect. However, it will be even more interesting if this small change in the choice architecture has an effect on decision making.

3.1.3 Part three: Background questions

The third and last part is questions about the participants' background. This part includes questions about gender, age, geography, living area, ethnicity, education, economic education, occupation, income, marital status and number of children. In addition, there is a question about time preferences. As discussed in the literature review, Heijden et al., (2011) find that impatient individuals are more influenced by a nudge than patient ones. We find it interesting to test whether this personality trait also is related to the default effect.

We chose to measure time preference in a slightly different way than in this study. We base our measurement on Dohmen et al. (2015), which makes use of the question "How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?" This question is general and does not only apply to the financial context specifically. We use a seven-point Likert's scale where 1 means you are "completely unwilling to do so" and 7 means you are "very willing to do so".

The control variables are important because they allow us to compare people with the same background. They eliminate rivalry explanations, enabling us to identify the relationship between the independent and the dependent variable. We want to control that the effect on the dependent variable is actually caused by the independent variables and not by a third variable, like education or ethnicity. This will increase the probability of revealing a true causal relationship.

3.1.4 Overview of experiment

One of the main benefits of our design is that we have control over the experimental situation and context. For this reason we can eliminate most other causes to the variation in the dependent variable, in our case risk-taking behavior. An example of a rivalry explanation is different perceptions of the probability of winning the lottery. If men perceive this probability differently than females, this could cause a difference in risk behavior. We want to isolate the effect of the default, and thus eliminate this possibility by providing the participants with the probabilities of winning and losing the lottery.

If the choice between safe payment and lottery was to take place in a real life situation, factors such as other incentives and different pay-off structures could cause variation. This is the drawback of a field study. Even though field studies better capture reality and real life choices than experimental studies, it may be difficult or impossible to isolate specific causes of the observed effects. In experimental studies, we are permitted to change each variable in a controlled and systematic manner. In our study, we place men and women in the same situation and the same context, thereby isolating the cause of variation in risk behavior to the variation in default bias and gender.

The experiment is outlined in Table 1 below.

Experiment Groups	Stimuli	Output
Risky Default Group	Risky Option as Default (X_R)	Male Average Risk-taking (X_{RM}) Female Average Risk-taking (X_{SF})
Safe Default Group	Safe Option as Default (X_S)	Male Average Risk-taking (X_{RM}) Female Average Risk-taking (X_{SF})

Table 1: Overview of the two experimental groups

There are two effects influencing risk taking. The first one is the treatment effect, measured as the difference between average risk-taking in the Risky Default Group and average risk-taking in the Safe Default Group, $(X_R - X_S)$. If the default has an impact on people's choices, we would see that participants initially getting the lottery ticket as payment would be more likely to keep the lottery, and not exchange it for the safe payment. Participants initially getting the safe payment, would be more likely to keep the safe payment, and not exchange it for the lottery. In this case, X_R would be larger than X_S, making the treatment effect $(X_R - X_S)$

positive. If the default has no effect on risk-taking behavior, X_R and X_S would be equal, making $(X_R - X_S) = 0$.

The second effect is the gender effect, measured as the difference between average risk-taking behavior among men and average risk-taking behavior among women, $(X_M - X_F)$. If males are more risk seeking than females, X_M will be larger than X_F, making the gender gap positive, $(X_M - X_F) > 0$. If females are more risk seeking than males, the gender gap will be negative, $(X_M - X_F) < 0$. If we find no difference in risk-taking behavior among males and females, there will be no gender gap, $(X_M - X_F) = 0$.

The effect we are interested in is the product of these two effects together, namely the interaction effect of gender and treatment. Instead of looking at the treatment effect on risk-taking behavior in general, we want to look at the treatment effect on the gender gap in risk-taking behavior. We intend to investigate whether females are more influenced by the default bias than males. If this is the case, the females in the Risky Default Group will be more influenced by the risky default than males in this group. This will make females come closer to the risk-taking behavior of males, which in turn will decrease the gender gap $(X_{RM} - X_{RF})$. Under the same assumption, the opposite will happen in the Safe Default Group. Here, females will be more affected than males by the safe default, making females even more risk averse on average compared to males. Thus increasing the gender gap $(X_{SM} - X_{SF})$.

Our hypothesis is that we will find evidence of a positive interaction effect, where the gender gap in the Safe Default Group is larger than the gender gap in the Risky Default Group. This can be mathematically formulated as: $(X_{SM} - X_{SF}) - (X_{RM} - X_{RF}) > 0$.

3.2 Conducting the experiment

This subchapter explains how our experiment was executed and consequently how our data was collected. The subchapter starts with a short discussion on online experiments and a presentation of the platform used to recruit participants. We have performed power calculations to determine the required sample size for our experiment, which will be presented in section two. The third section explains how the experiment was executed, and describes some of the most important considerations we did regarding the implementation. The last section provides a brief overview of the sample. A more comprehensive documentation of the

details regarding the experiment is provided in section A.2 in the Appendix. Ethical considerations regarding our study are evaluated in Appendix A.3.

3.2.1 Online Experiments and Amazon Mechanical Turk

We made use of an online experiment instead of a lab experiment. Online experiments have recently become very popular, and have several advantages compared to the traditional lab experiment. One of the most important advantages is that online experiments are easier to conduct. A lot of the work that is done manually in lab experiments, is done automatically in online experiments (Dandurand, Schultz and Onishi, 2008). Automation further increases flexibility and saves time and resources. It is a relatively inexpensive way of reaching a large and more diverse sample compared to the standard student population often used in lab or field experiments (Rademacher and Lippke, 2007).

Furthermore, it is easier and faster to get enough participants and to get the right kind of sample. This is important to be able to generalize the results to wider populations[3]. Online experiments permits the participants to conduct the experiments in the comfort of their own home, which may serve as a more natural decision-making environment than a lab. This might also cause less stress on the participants (Duersch, Oechssler and Schipper, 2009; Vinogradov and Shadrina, 2013). Research directly comparing results from experiments conducted online and in a laboratory setting have generally found consistent results, especially for shorter and simpler experiments (e.g. Dandurand et al., 2008; Gosling, Vazire, Srivastava and John, 2004; Meyerson and Tryon, 2003; Riva, Teruzzi and Anolli, 2003).

The participants for our study are recruited through the platform Amazon Mechanical Turk (mTurk). mTurk is an online global marketplace created by Amazon. On this platform workers choose which jobs or Human Intelligent Tasks (HITs) to do for pay. It is used by a growing body of researchers to conduct economic experiments. mTurk has one of the largest subject pools available among crowdsourcing[4] platforms (Mason and Suri, 2011). Hence, mTurk is convenient and enable us to collect data from a large and diverse subject pool, at a low cost, in a short amount of time, reducing geographical and financial constraints on research (Mason and Suri, 2011; Paolacci, 2012).

[3] See section A.6.1.2 about external validity for a further discussion about generalizability.
[4] Howe (2006, ref. in Mason and Suri 2012) defines crowdsourcing as "a job outsourced to an undefined group of people in the form of an open call".

Several researchers have investigated the representativeness of mTurk as a sampling frame. Paolacci (2012) directly compares mTurk participants to traditional subject pools. He finds consistent results with previous decision making research implying that a sample obtained from mTurk is as least as fit as traditional samples to draw general conclusion about tasks involving money and risk.

When it comes to preferences for time and money, it is worth noting that Paolacci (2012) finds mTurkers to be less extraverted, less emotionally stable, and to have lower self-esteem. In addition, Paloacci (2012) finds the attention levels to be lower among mTurkers than the other samples. We thereby have to acknowledge that mTurkers might be different from non-mTurkers on social and financial traits (Paolacci, 2012). However, "there are numerous studies that show correspondence between the behavior of workers on Mechanical Turk and behavior offline or in other online contexts. While there are clearly differences between Mechanical Turk and offline contexts, evidence that Mechanical Turk is a valid means of collecting data is consistent and continues to accumulate" (Mason and Suri, 2004, p. 4). A more extensive discussion of the benefits and weaknesses of using this platform is provided in the sections on validity and reliability in Appendix A.6.

We chose to narrow our sample to include only Americans. The sample was not limited with regard to any other background variables. This was done because we wanted the sample to mimic the US population. In addition, we wanted a diverse and rich sample to be able to analyze the impact on risk-taking of different backgrounds. By using mTurk as a sampling frame, the sample can be categorized as a non-probability sample with a self-selection sampling technique. Each participant has to decide for themselves if they want to be a part of our study. This is not the optimal sampling method for a causal study (Saunders, Lewis and Thornhill, 2009). A better method would have been to randomly select the desired number of participants from the total population, in our case the total US population. As this is not possible, the self-selection sampling method is the best feasible option in our case.

3.2.2 Power calculations

Power calculations (or power analysis) is a process for determining the sample size for a research study. The sample size depends on the desired level of statistical significance, statistical power and the expected effect size (Cohen, 1992). Statistical power could be

explained as the probability of determining a *true* effect when it exists. The input variables will vary from study to study. In most cases, power analysis involves a number of simplifying assumptions. Consequently, a power calculation will not give a 100 % correct answer.

We made use of the statistical software GPower to analyze the needed sample size. Cohen (1992) recommends a standard significance level (α) of 0.05 and a power (1-β) of 0.80. We will use both t-tests and multiple linear regression in our analysis. For a one-tailed t-test the total required sample size is 620 (310 in each group) to detect small effects, 102 (51 in each group) to detect medium sized effects, and 42 (21 in each group) to detect large effects. For a two-tailed t-test the required sample sizes are 788, 128 and 52 respectively, for small, medium and large effects.

To determine the required sample size for the regression analysis, we will again use a significance level (α) of 0.05 and a power (1-β) of 0.80. The number of tested predictors in our regression is 3 (gender, treatment and gender*treatment), while the total number of predictors (which includes control variables) is 12. The required sample size is 550 to detect small effects, 78 to detect medium effects and 35 to detect large effects. The figure below shows the required sample size on the y-axis and the effect size on the x-axis. For multiple linear regressions, one usually regard effect sizes of about 0.02 to be a small effect and effect sizes of about 0.15 to be a medium effect. The blue line implies a power (1-β) of 0.90, and the red line a power (1-β) of 0.80. Thus a higher power requires a larger sample size.

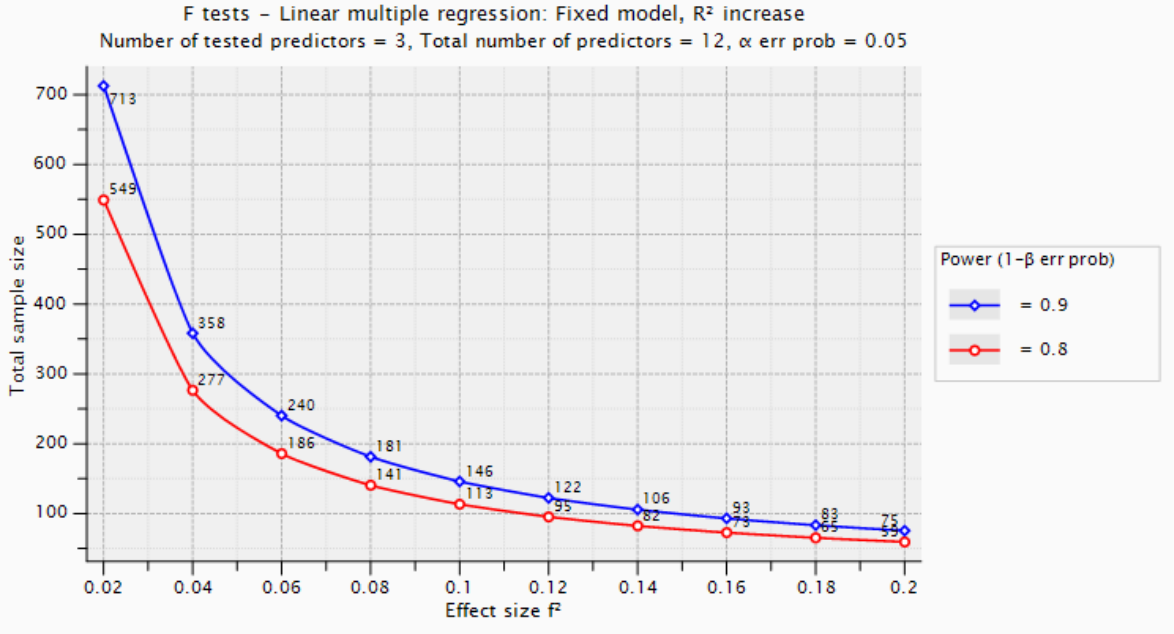


Figure 3: Required sample size for multiple linear regression

We could not find previous studies investigating the same effect as we are interested in (the interaction between gender and default effect), which also found significant results. Consequently, we cannot be sure about what effect size to expect. The main effect in focus (the interaction effect) is a so called *difference in difference* (difference in treatment effect in the difference between the genders), which further complicates the calculation. If we were to only investigate the gender difference, which is a simple difference, we would have previous studies to look to. For instance Hartog, Ferrer-i-Carbonell and Jonker (2002), which conducted three separate surveys with large sample sizes (2011, 1599 and 17097 respondents). They estimated a risk aversion parameter, and found women's estimated parameter to be 10 % to 30 % larger than men's. This effect size is generally categorized as a small effect for t-tests. Therefore, we would need a sample size of up to 620 to detect a similar effect with a one-tailed t-test.

Since we are doing an incentivized experiment, our sample size is restricted by budget constraints. The Choice Lab[5] helped us finance our study, and we developed a sensible budget and sample size in collaboration with our supervisor Alexander W. Cappelen. The total sample size of our study is therefore 360 participants. With an α of 0.05 and a power (1-β) of 0.80, this sample size would enable us to detect an effect size of 0.03 in our regression analysis (in line with the graph above). These numbers seem promising, and we believe a sample size of 360 will be sufficient to detect potential effects.

3.2.3 Implementation and execution

After the experimental design had been developed, applied and approved, the experiment was ready to be launched on mTurk. It was carried out in collaboration with Ph.D. students at The Choice Lab, who were more experienced with experiments on mTurk. They made sure that there were no deficiencies in our design and that the survey ran smoothly. This was ensured by running the survey in a test environment on mTurk called "Sandbox". Here one can make sure that the connection between mTurk and our survey in Qualtrics was working fine.

The qualification and "quality" of the participants was ensured through specifications in the survey. When the survey was launched, we were able to set criterions that the mTurk workers were required to meet in order to work on our HIT. We required the worker's HIT Approval

[5] The Choice Lab is a research group at the Department of Economics, at the Norwegian School of Economics.

rate[6] to be 95 % or greater. In addition, the workers must previously have taken 1000 or more HITs to be able take our survey. These actions ensure that the participants are familiar with the process of conducting HITs and surveys.

Ideally, we would want the number of males and females in our sample to be identical. However, there are no feature on mTurk to ensure this. An alternative would be to use a method developed by researchers at The Choice Lab. This method gives an equal amount of male and female participants by first allowing females to take the HIT. When the desired amount of female partisipants is reached (in our case 180), the HIT is closed for females, and opened for males. When the same amount of male partisipants is reached, the HIT is closed.

There are several drawbacks with this technique. Firstly, workers on mTurk are not automatically registrerd with gender. Thus, participants who previously has conducted experiments carried out by The Choice Lab, are registered as males or females. When using the method of equal gender composition, the only workers who can participate in the study are those registered by The Choice Lab. This may prove problematic as it might cause selection bias. The sample is further affected because all females are recruited at a different point in time than men. As an example, if we assume that the study got all its 360 responses during 10-12 hours, it might happen that all females take the survey in the morning and midday, while all males take the survey in the evening. This might cause biased results.

A third drawback is related to the fact that the pool of participants is significantly restricted when using this method. When the only workers allowed to do our survey are those registered by The Choice Lab, it would take considerably longer to get the desired sample size. It could take several days instead of a few hours. In summary, the probability of something going wrong is considerably larger when using this method compared to executing a "normal" HIT where everyone (who meets the requirements) are allowed to participate. Based on an evaluation of benefits and weaknesses we decided to use a normal HIT, despite the risk of getting an unequal distribution of males and females. The decision was justified by the fact that the distribution of male and female workers on mTurk is rather equal, although there are slightly more males than females.

[6] The HIT Approval Rate is a "System Qualification". "A Worker's HIT Approval Rate is calculated as the Workers' Lifetime Assignments Approved divided by the Worker's Lifetime Number of Assignments Submitted - on ALL Assignments" (The Mechanical Turk Blog, 2011).

3.2.4 Sample

Our planned sample size was 360 participants. This was the number of participants completing the survey in Qualtrics. However, a technical failure caused 371 participants to be registered on mTurk, and only 354 to be included in our dataset. This implies that our sample size ended up being 354 instead of 360. As this is such a small deviation (0.017 %), it will not influence our data analysis or results. In chapter A.1 in the Appendix is an overview of the descriptive statistics of our sample based on background characteristics.

The sample consists of 55 % males and 45 % females. The distribution of participants according to occupational status, ethnicity and education is presented below in Figure 5.In our distribution, most participants belong to the occupational group Employed (62 %). Furthermore, the largest ethnic group by far is White Americans, accounting for 80 % of the total sample, followed by Asian Americans (5.9 %) and Black or African Americans (5.6 %). In terms of education, the distribution is more even, with most participants having 4-year college (38 %) or some college education (27 %). In addition, 61 % of participants have no education in economics.

To assess whether the sample is representative for the US population, we made use of statistics provided by the United States Census Bureau, which is part of the U.S. Department of Commerce and provides high-quality economic analysis. The statistics are from 2013 and 2014, but it is reasonable to believe that the statistics have not changed much in 1-2 years. The average age of our sample is 35.0 years, and the median age is 33 years. The reported median age of the U.S. population is 39.0 (“U.S Census Bureau”, 2014a). The gender distribution is 49.2 % male and 50.8 % female (“U.S. Census Bureau”, 2014b). 50.6 % of the population is married (“U.S Census Bureau”, 2013a), compared to 40 % in our sample.

Concerning education, only 1 % of our sample has completed less than High School. In the general population this figure is 13.9 % (“U.S Census Bureau”, 2013a). The proportion who has two years College or more is 58 % in our sample and 37 % in general (“U.S Census Bureau”, 2013a). Concerning ethnicity in the U.S. today, 63.3 % are White Americans, 16.6 % are Latin or Hispanic, 12.2 % are Black, and 4.8 % are Asian (“U.S Census Bureau”, 2013b). Consequently, our sample is younger, more male dominated, higher educated and includes less ethnic minorities than the general U.S. population. In other words, the sample is rather similar, although not perfectly representative.

Figure 4: Descriptive statistics of the sample

4. Results and analysis

In this chapter we attempt to answer our research question by analyzing the data we have gathered in our experiment. First, we present some descriptive statistics to reveal the direction of our results and an initial idea of our findings. We make use of t-tests to investigate the results further and uncover whether any of the findings are statistically significant. As these tests cannot be used on interaction variables and do not include control variables, we perform multiple linear regressions to analyze the results while controlling for background variables. Lastly in this chapter, we do a short additional analysis to investigate the relationship between the default effect and people's time preferences.

4.1 Main analysis: Risk-taking behavior based on gender and treatment

In the analysis, risk-taking behavior is measured as the percentage of participants choosing the risky option (lottery). Regarding the treatment effect, we find that the average risk-taking in the Risky Default Group is 35.79 %, while the average risk-taking in the Safe Default Group 31.46 %. This infers that the magnitude of the overall treatment effect for both genders, $(X_1 - X_2)$, is 4.33 %, implying an increase of 11.24 %. This is the increase in risk-taking behavior from the Safe Default Group to the Risky Default Group. Considering the gender effect, we find that the overall average risk-taking among males is 35.90 %, while the same numbers for females is 30.82 %. Hence, the overall gender effect for both treatment groups, $(X_M - X_F)$ is 5.08 %, which is equivalent to an increase in risk-taking of 28.88 %.

The figure below presents our main findings divided by both gender and treatment. It shows the average risk-taking behavior of males and females in the two treatment groups. In the Risky Default Group 35.79 % of males chose the risky payment and 35.80 % of females chose the risky payment. In the Safe Default Group 36.00 % of males chose lottery, while 26.64 % of females chose lottery. What is interesting about these results is that when splitting the average risk-taking on both gender and treatment, we find that the treatment effect is non-existing among males, and thus even more apparent among females. For females the treatment is increasing the average risk-taking by 10.61 %, which implies an increase of 34.38 %[7]. This is a large effect, as the risky treatment increases the proportion of women choosing the risky option by one third.

[7] The increase from X_{2F} to X_{1F} is: (35.80 % / 26.64 %) – 1 = 34.38 %

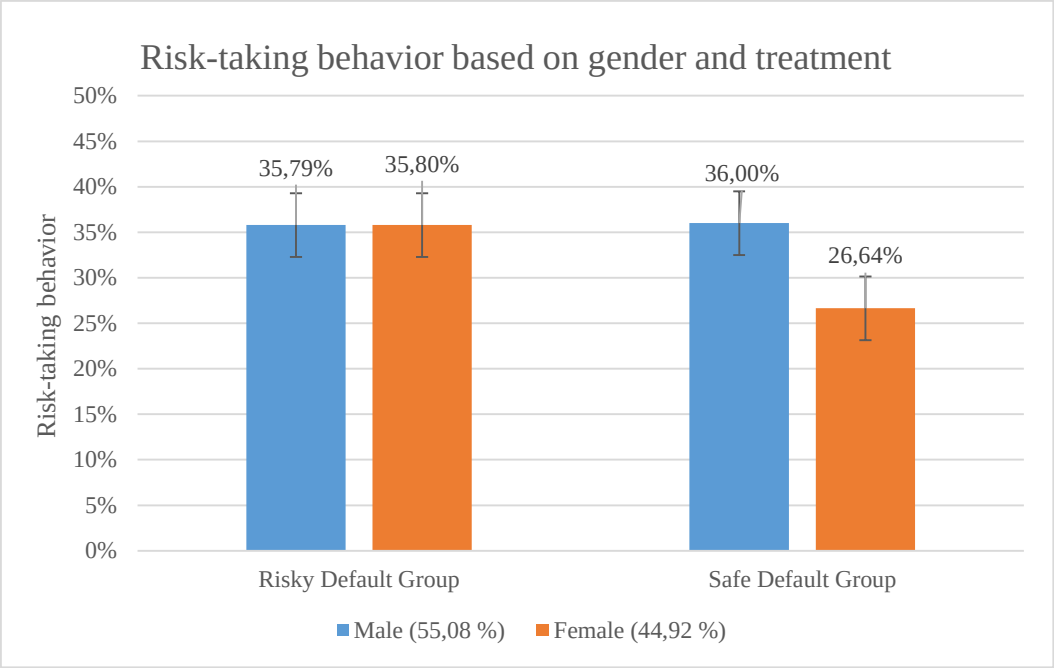


Figure 5: Graphical overview of main results

The gender effect differs substantially between the two treatment groups. In the Risky Default Group there is no difference in risk-taking behavior among males and females. In other words, when the default is set to a risky option, the gender gap disappears and females take equal amount of risk as males. On the other hand, in the Safe Default Group, the difference in risk-taking behavior is 10.36 %. This represents an increase of 35.13 % from females to males, which is a substantial effect size. The results appear to support our expectations of a gender difference, a treatment effect and a gender difference in the treatment effect. If the differences are large enough to be statistically significant needs to be tested.

We want to test whether the treatment effect and the gender effect is significantly different form zero. This is done by comparing the average risk seeking behavior of the different subgroups in our sample. If the difference between the averages is large enough, the result will be statistically significant, implying that we can draw a confident conclusion on a particular significant level. All tests are two-sample t-tests of means with equal variances. Because we initially were unsure about the direction of our results, and whether they would be in line with our hypothesis, we conducted both one-tailed and two-tailed tests.

When testing the treatment effect, the null hypotheses is that there is no difference between the Risky Default Group and the Safe Default Group in their average risk taking. The

alternative hypothesis for the two-tailed test is that the average risk-taking in the Risky Default Group is different from that of the Safe Default Group. For the one-tailed test, the alternative hypothesis is that the average risk-taking in the Risky Default Group is greater than that of the Safe Default Group. For males, the p-value of the two-tailed t-test is 0.98, implying that there is no evidence to reject the null hypothesis. The same results are seen in the one-tailed t-test, with a p-value of 0.49. In other words, this test shows that for males there is no evidence of a treatment effect, and consequently no evidence of a default bias.

Considering the treatment effect among females, the p-value of the two-tailed t-test of 0.17, indicating that there are some, although not sufficient evidence to reject the null hypothesis. For the one-tailed test, the p-value is half the p-value of the two-tailed test, namely 0.08. Consequently, the t-test proves that for females the average risk-taking in the Risky Group is greater than in the Safe Group, significant at the 90 % level. Thus, there is evidence of a treatment effect, or default bias, among females.

The null hypothesis of the t-test to test the gender effect is that there is no difference between males and females in their average risk taking. The alternative hypothesis for the two-tailed test is that the average risk-taking among males is different from that of females. For the one-tailed test, the alternative hypothesis is that the average risk-taking for males is greater than that of females. For the Safe Default Group, the p-value of the two-tailed t-test is 0.14 and the p-value of the one-tailed is 0.07. The one-tailed t-test proves that there is enough evidence to reject the null hypothesis at the 90 % significance level. This means that there is a gender gap in risk-taking behavior in the Safe Default Group with males being more risk seeking than females. For the Risky Default Group, the p-values for the two-tailed and one-tailed tests are 0.99 and 0.49 respectively. Thus, there is no evidence to reject the null hypothesis. In other words, the gender effect is only apparent when the participants have received the safe default and not the risky default.

The effect we are most interested in is the interaction effect between treatment and gender. To investigate this effect further we make use of multiple linear regression. This technique allows us to identify the effects of gender and treatment on risk-taking behavior, while controlling for background variables. The general form of our regression equation is as follows:

Risk-taking = β_0 + β_1* RD + β_2* F + β_3*RD*F + β_x* Control variables + … + ε

RD (Risky Default) is an indicator variable for treatment group, taking the value 1 if the participant is in the Risky Default Group, and 0 if not. F (Female) is an indicator variable for gender, taking the value 1 if the participant is female, and 0 if male. RD*F is an interaction variable for treatment group and gender. This variable takes the value 1 if the participant is female and is in the risky default treatment group, if not, the value is 0. The reference group (base group) is males in the safe default treatment group.

The dependent variable is risk taking, and is measured as the percentage of participants choosing lottery. If all participants chose lottery, the variable will be 1. If all participants chose safe payment, the variable will be 0. In the table below, the output of our main regressions is presented. Regression (1) includes only the indicator variable for treatment group and shows that the average treatment effect on risk-taking is 4.33%. This is the average increase in risk-taking behavior from the Safe Default Group to the Risky Default Group. Regression (2) includes the indicator variables for both treatment and gender, and shows that the size of the gender effect is 5.18 %. In addition to treatment and gender, regression (3) includes the interaction variable between gender and treatment, which in this regression is 10.4 %. The regression shows that the average risk-taking in the safe default group is 36.0 % among males and 26.5 % among females. In the risky default group average risk-taking is 35.79 % among males and females. These numbers are consistent with the averages computed in the beginning of this chapter.

Regressions (4)-(6) control for background variables. The background variables included are age, income, economic education and time preference. These specific variables are included because they are the most significant ones when controlling for all background variables. We excluded variables having very little explanatory power, thus being redundant in the analysis. The control variables did not yield large difference in the beta value of the gender effect, compared to the regression with no controls. The treatment effect is reduced, but it is still not significantly different from zero, and thus not different from the regressions without control variables. This consistency was expected as the two treatment groups were randomized.

Table 2: Regression analysis: Effect of treatment and gender on risk-taking

	(1)	(2)	(3)	(4)	(5)	(6)
Risky Default	0.0433	0.0445	-0.00211	0.0210	0.0226	-0.0497
	(0.0503)	(0.0503)	(0.0678)	(0.0513)	(0.0512)	(0.0678)
Female		-0.0518	-0.104		-0.0563	-0.135**
		(0.0506)	(0.0715)		(0.0509)	(0.0681)
Female×Risky Default			0.104			0.159
			(0.101)			(0.0991)
Constant	0.315***	0.337***	0.360***	0.463***	0.493***	0.530***
	(0.0355)	(0.0418)	(0.0473)	(0.0732)	(0.0783)	(0.0802)
Background variables	No	No	No	Yes	Yes	Yes
Linear combination of RD and Female×RD			0.102			0.109
			(0.075)			(0.074)
Observations	354	354	354	354	354	354

Note: The table reports linear regressions of the variable "Risk taking".
"Risky Default": indicator variable taking the value one if the participant is in the Risky Default treatment.
"Female": indicator variable taking the value one if the participant is female.
"Female × Risky Default": interaction between "Female" and "Risky Default".
Background variables are age (consisting of four indicator variables: Age below 26, age 27-30, age 31-35 and age above 36), economic education (consisting of three indicator variables: No economic education, 1year of economic education, 3 years or more of economic education), "Income below USD 50 000": indicator variable taking the value of one if the participant has a yearly average household income below USD 50 000, and "Impatient": indicator variable taking the value of one if the participant has a score in the lower half of the time preference scale (1-4 out of 7).
Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Regression (6) includes the indication variables for treatment and gender, the interaction variable between these two, and control variables. The constant (β_0) is 0.530, meaning that the average risk-taking for males in the Safe Default Group is 53.0 %. The coefficient for Risky Default (β_1) is -0.0497 and has a p-value of 0.46, meaning that it is not significantly different from zero. This implies that the average increase in risk seeking behavior among males in the risky default treatment group compared to the safe default treatment group is zero. In other words, males seem to be equally risk seeking irrespective of which treatment group they are in. The number of males in the Risky Default Group that is keeping the lottery is equal to the number of males in the Safe Default Group that is exchanging the safe payment for a lottery.

The coefficient for Female (β_2) is -0.135, indicating a gender difference in risk-taking of 13.5 %. This suggests that the average risk-taking for females in the safe default treatment group is 39.5 %, compared to 53.0 % for males. The p-value of this coefficient is 0.048, meaning that the gender gap in risk-taking is statistically significant at the 95 % level, for the Safe Default Group. The average increase in risk seeking behavior for females in the Risky Default Group compared to females in the Safe Default Group, is 10.9 %, which is the linear combination of β_1 and β_3. This figure has a p-value of 0.14 for a two-tailed test and 0.07 for a one-tailed test, implying that the treatment effect among females in statistically significant at the 90 % level.

The coefficient for Female × Risky Default (β_3) represents the interaction effect, which is the effect we are most interested in. The value of β_3 is 0.159, and because the value is different from zero, there is indications of an interaction effect. The p-value of this coefficient is 0.11, suggesting that the interaction effect is not statistically significant. However, because we have a hypothesis stating an expectation of a positive interaction effect, we can make use of a one-tailed t-test instead of a two-tailed t-test. The one-tailed t-test provides a p-value of 0.055. This infers that the interaction effect is statistically significant at the 90 % significance level. Thus, we have found evidence of a causal relationship between the gender difference in risk-taking behavior and the gender difference in treatment effect, i.e. default effect.

4.2 Additional analysis: time preference

In this subchapter, we briefly examine another aspect of our results. We test whether the default bias is related to time preferences. Being impatient means that you prefer having a benefit today, rather than waiting for it and having a greater benefit in the future. In an economic context, this implies that you have a higher discount rate. Figure 6 below shows the average risk-taking and reported time preference of our sample. Time preference is measured on a 7 point scale, with 1 being very impatient and 7 being very patient. In this analysis, time preference is converted into an indicator variable, denoted "Impatient". This variable takes the value 1 if the participant has a time preference of 1 to 4, and 0 if the participant has a time preference of 5 to 7. The figure displays large differences in risk-taking behavior, with impatient individuals being a lot more risk averse than patient ones.

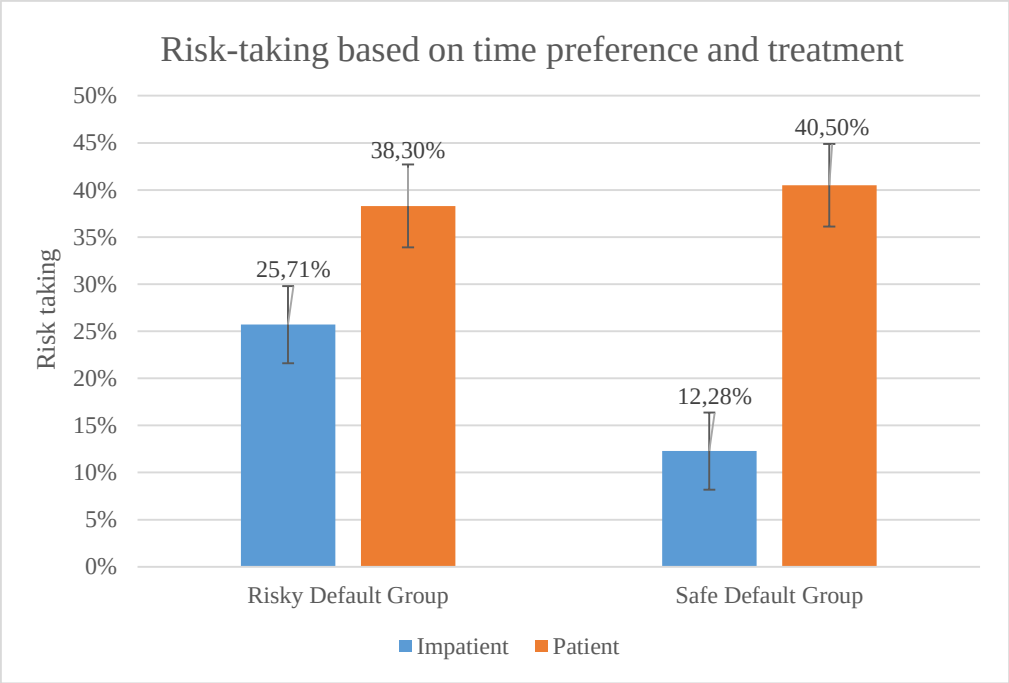


Figure 6: Default bias based on time preference

The average risk-taking among patient individuals is 38.30 % when exposed to a risky default, and 40.50 % when exposed to a safe default. Among these individuals, the difference between the two treatment groups is not significantly different from zero (p-value: 0.72). This suggests that patient individuals are not affected by the default effect. For the impatient individuals we find a different result. In the Safe Default Group, the average risk seeking is 12.28 %. In the Risky Default Group, this figure is more than doubled to 25.71 %. The treatment effect among the impatient individuals is almost statistically significant at the 90 % level (p-value: 0.101), when using a two-tailed t-test.

Table 3 below presents the output from the regression analysis. Regression (1) shows that impatient individuals are on average 21.9 % more risk averse than patient ones, significant at the 99 % level. Regression (2) suggests that the overall treatment effect is less than 2 %, implying that when participants are not divided by time preference there is no evidence of a default bias. Regression (3) includes the interaction variable of time preference and treatment. The linear combination of Impatient and Impatient × Risky Default proposes that impatient individuals are 12.6 % more risk seeking when exposed to a risky default compared to a safe default. This is consistent with the numbers from figure 6 above. The p-value for the interaction variable is 0.17 for a two sided t-test. As we expected to find results similar to

those of Heijden et al. (2011), we can make use of a one sided test. The p-value for this test is 0.09, signifying that the interaction effect is significant at the 90 % level.

Table 3: Regression analysis: Effect of time preference and treatment on risk taking

	(1)	(2)	(3)	(4)	(5)	(6)
Impatient	-0.219***	-0.217***	-0.282***	-0.215***	-0.212***	-0.277***
	(0.0562)	(0.0568)	(0.0745)	(0.0502)	(0.0514)	(0.0637)
Risky Default		0.0171	-0.0220		0.0210	-0.0174
		(0.0498)	(0.0575)		(0.0513)	(0.0612)
Impatient×Risky			0.156			0.155
			(0.115)			(0.106)
Constant	0.393***	0.384***	0.405***	0.474***	0.463***	0.481***
	(0.0287)	(0.0393)	(0.0422)	(0.0682)	(0.0732)	(0.0749)
Background var.	No	No	No	Yes	Yes	Yes
Linear combination of Imp. and Imp×Risky			-0.126			-0.122
			(0. 088)			(0. 085)
Observations	354	354	354	354	354	354

Note: The table reports linear regressions of the variable "Risk taking".
"Impatient": indicator variable taking the value of one if the participant has a score in the lower half of the time preference scale (i.e. a score of 1-4 out of 7).
"Risky Default": indicator variable taking the value one if the participant is in the Risky Default treatment.
"Impatient × Risky": interaction between "Female" and "Risky Default".
Background variables are age (consisting of four indicator variables: Age below 26, age 27-30, age 31-35 and age above 36), economic education (consisting of three indicator variables: No economic education, 1year of economic education, 3 years or more of economic education), and "Income below USD 50 000": indicator variable taking the value of one if the participant has a yearly average household income below USD 50 000. Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Regressions (4)-(6) provides the same output as regressions (1)-(3), while controlling for the relevant background variables. These regressions suggest that the findings are not altered when control variables are included. The size of the interaction effect is 15.5 % and the p-value is reduced to 0.14 for a two-sided test, and 0.07 for a one-sided test. This suggests that there is evidence of impatient individuals being more biased by a default effect, significant at the 90 % level. These findings expand the implications of the study by Heijden et al. (2011), inferring that impatient individuals not only are more influenced by the nudge used in their study, but also by a default nudge, or default effect.

In addition to time preference, we looked into the other background variables, to see if we could find indications of who is most affected by the default effect. In Chapter A.3 in the appendix, we provide an overview of the change in risk-taking between the Safe Default Group and the Risky Default Group, based on the background variables. However, when the sample is divided in several subgroups, the sample size in each category becomes rather small, increasing the risk of unreliable results. We found no results to be significant at the same time as the sample size was sufficiently large enough to be reliable.

5. Discussion and conclusion

In this study we found evidence of a causal relationship between the gender difference in risk-taking and the gender difference in default bias. We investigated two separate effects and the interaction between them, thereby combining two different literatures. The first effect is the gender effect on risk taking. In accordance with our hypothesis, we found indications of a gender gap in risk-taking behavior with females being 28.88 % more risk averse than males (26.94 % with control variables). The overall gender effect for both treatment groups is statistically significant at the 90 % level when using a one-sided t-test (p-value: 0.07). When including control variables, the effect is significant at the 95 % level (p-value: 0.04).

Furthermore, the direction of our result indicate evidence of a treatment effect, i.e. a default bias. Even though the two treatment groups where exposed to identical choices, with only a small change in the question structure, this little twist influenced their behavior. According to standard economic theory, this should not have occurred, as people should behave rational and consistent regardless of the framing of choices. We found participants to be 11.24 % more risk seeking in the Risky Default Group than in the Safe Default Group. However, the regression analysis reveals that the overall treatment effect (for both genders) is not statistically significant at the 90 % level.

Our most interesting result is found when interacting this treatment effect with gender. When dividing our sample by gender, we find a significant treatment effect among females, and no evidence of a treatment effect among males. This implies that only females seems to be affected by the default bias, while males are equally risk seeking irrespective of the default they receive. Moreover, this creates a larger gender gap in risk-taking behavior in the Safe Default Group, while in the Risky Default Group the gender gap is not only reduced, but completely non-existing. In the latter group, males and females are equally risk seeking. In summary, we find that it is possible to influence the gender gap in risk-taking behavior. By framing the choice as having a risky default, the gender gap disappears. In other words, we found that the choice of default has a causal effect on the gender gap in risk-taking behavior.

In addition to our main analysis, we did a more comprehensive investigation on one of the background variables. We examined the relationship between time preference and default

bias. Our findings prove that impatient individuals are more influenced by the default effect than patient ones, significant at the 90 % level. These results are consistent with our expectations and the research of Heijden et al. (2011), who also finds that the decision frames of impatient people are affected more easily than those of patient people. This is interesting from a policy perspective because people who often are associated with undesirable behaviors such as low savings, overspending on credit cards and obesity, are the same people who are mostly affected by the default effect. These people are also the target group of many nudges, and our findings thus confirms that the target group are those who actually are most influences by the default nudge.

In relation to existing literature, our study both confirms and expands prior knowledge. The literature on risk-taking behavior is extensive. We found opposing arguments regarding the gender gap, depending on the framing of the studies, the control variables included, and other factors influencing the results. However, there seem to be more evidence in favor of females being more risk averse than males (e.g. Byrnes et al., 1999; Charness and Gneezy, 2011). This conclusion is supported by our results, as we also find females to be more risk averse than males.

Regarding default bias, there is very little previous research on gender differences. We could only find one study touching upon this question, and they found little evidence of a gender differences or a default bias at all. Hence, our study provides a contribution to this field. Many studies on default options provide evidence that the default effect influence people's choices. Therefore, we expected that both genders would be affected by this bias, although females would be more influenced. This is not what our results suggest. We find evidence of a default bias among women, although no such evidence was found among males. In other words, we found a great gender difference in this bias.

In summary, our findings are largely in compliance with our hypothesis, and because our hypothesis is based on results from previous research, this implies that our results are consistent with existing literature. We argue that our finding regarding the gender gap in risk-taking is confirming prior research, while our finding regarding the gender gap in default bias is expanding prior research. We have contributed to filling a gap in the literature in the intersection between gender differences in risk-taking and default bias.

The implications of our study are mainly related to framing of questions. We have demonstrated that using a default option as part of the framing will affect the choice of females, while men appear unaffected. If it is desirable to make women less risk averse, it will be important to consider the framing of the choices. For women, it would be beneficial to be aware that one might be affected by defaults and that our choice might be influenced by the choice architecture and even small changes in the context. If people are attentive, each individual can make a more cautious choice by deliberately evaluating whether it is beneficial to go with the default (the recommendation) or to make your own assessment of the option set. Actors using defaults should be aware that it might influence different people differently, and cause results that differs from their intention. This is particularly important for governments and policy makers.

Two alternative hypotheses to why women are more risk averse can be proposed. Firstly, that women have different underlying risk preferences than males. Secondly, that choices regarding risk often is framed in a setting where the safe option is the default. In this case, the gender difference is driven by a greater default bias among females, rather than a greater risk aversion. As our findings indicate, the gender difference disappears when the question is presented with a risky option as default. With our research design we cannot answer which of these hypotheses that are the true one. We end this discussion by stating that more research is needed in this field to reveal the true links and mechanisms behind the possibilities of affecting people's risk-taking behavior. Limitations of our study and further suggestions for future research will be discussed in the next subchapter.

5.1 Limitations and suggestions for future research

The findings and limitations of our study provide implications for future research. One limitation of our study is that using incentivized choices and real money is costly and difficult to perform with a large, representative sample (Dohmen et al., 2009). The financial restriction of our research implies a tradeoff between substantial incentives and a sufficient sample size. As a result, we decided to conduct our experiment with low stakes to be able to obtain a satisfactory sample size, at the same time as we adhere to our financial boundaries. Thus, we were able to attain an adequate and satisfactory sample size of 354 participants.

One potential problem is that experiments with small stakes might not yield conclusions that generalize to high stake experiments. In this regard it has been raised concerns about the quality of work on mTurk due to the low payment. Mason and Suri (2011) states that mTurk can be seen as a market for lemons. "However, there is usually a change in behavior going from paying zero to some low amount and little to no change in going from a low amount to a higher amount. Thus, the norm on mTurk of paying less than one would typically pay laboratory subjects should not impact large classes of experiments" (Mason and Suri, 2011, p. 9). Other studies have also shown that conclusions from lower stake experiments do generalize. As an example, the gender difference in risk-taking is found to be consistent across high stake and low stake experiments (Croson and Gneezy, 2009).

Even though we obtained a satisfactory sample size, the power of our analysis would have been improved if we had a lager sample. Ideally the sample would mimic the US population, but workers on mTurk are younger and higher educated than the general US population, and this was also reflected in our sample. Yet our sample is at least as representative of the population (in our case the US population) as samples drawn from traditional subject pools. In Appendix A.6 about validity and reliability, strengths and weaknesses of our study is further discussed and evaluated. In summary, we find the internal validity to be a strength of our study, while the external validity often is perceived as a threat to experimental studies.

The experimental setting is often stated to be artificial or non-realistic (e.g. Mook, 1983). To test the robustness and generalizability of our findings there is a need for studies investigating the topic further in other contexts (including field experiments), larger samples as well as for other populations. In relation to this, it would be interesting to investigate whether a person's true underlying risk preferences can be influenced or whether it is only the more shallow risk-taking behavior that is affected. Will risk-taking behavior adjust back to the person's true risk preference after the effect of a treatment fades, or would it be possible to affect risk preferences permanently? Is there a difference among males and females in this regard? And are there only certain stages of a person's life personality traits like these can be shaped?

Our study leaves many important questions unaddressed, but it provides significant evidence of the importance of framing on decision making. Providing a simple default option can alter people's behavior, and the insight from our study suggests that the gender gap in risk-taking can be influenced and even eliminated. Hopefully, our finding will create an interest in this question and encourage further research on the topic.

References

Agnew, J. R., Anderson, L. R., Gerlach, J. R., and Szykman, L. R., 2008. Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults. *The American Economic Review*, pp. 418-422.

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., and Tungodden, B., 2012. Willingness to compete in a gender equal society. Discussion paper, *Department of Economics*. Norwegian School of Economics, (24).

Amazon Mechanincal Turk, 2015. *Mechanical Turk is a marketplace for work*. [Online] Available at: https://www.mturk.com/mturk/welcome [Accessed 13 Oct 2015].

American Psychological Association, 2015. *Ethics Code Updates to the Publication Manual*. [Online] Available at: http://www.apa.org/ethics/code/manual-updates.aspx [Accessed 26 Oct 2015].

Arrow, K. J., 1956. Aspects of the Theory of Risk Bearing. Helsinki: Yrjo Jahnssonis Saatio.

Brinig, M. F., 1995. Does Mediation Systematically Disadvantage Women? *William and Mary Journal of Women and the Law*, 2, pp. 1-34.

Brown, C. L., and Krishna, A., 2004. The skeptical shopper: A metacognitive account for the effects of default options on choice. *Journal of Consumer Research*, 31(3), pp. 529-539.

Bryant, S.M., Hunton, J.E., and Stone, D.N., 2004. Internet-Based Experiments: Prospects and Possibilities for Behavioral Accounting Research. *Behavioral Research in Accounting*, 16, pp. 107-129.

Bryman, A., and Cramer, D., 2009. Quantitative Data Analysis with SPSS 14,15 & 16: A *Guide for Social Scientists*, London: Routledge.

Byrnes, J.P., Miller, D.C. and Schafer, W.D., 1999. Gender Differences in Risk Taking: A Meta-analysis. *Psychological Bulletin*, 125(3), pp. 367-383.

Charness, G. and Gneezy, U., 2011. Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior & Organization*, 83, pp. 50-58

Cohen, J., 1992. Statistical Power Analysis. *Current Directions in Psychological Science*, 1(3), pp. 98-101.

Cronqvist, H., and Thaler, R. H., 2004. Design choices in privatized social-security systems: Learning from the Swedish experience. *American Economic Review*, pp. 424-428.

Croson, R. and Gneezy, U., 2009. Gender differences in preferences. *Journal of Economic Literature*, 47(2), pp. 448-474.

Dandurand, F., Schultz, T.R., and Onishi, K.H., 2008. Comparing online and lab methods in a problem-solving Experiment. *Behavior Research Methods*, 40(2), pp. 428-434.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3), pp. 522-550.

Duersch. P., Oechssler. J., and Schipper, B.C., 2009. Incentives for subjects in internet experiments. *Economics Letters*, 105, pp. 120-122.

Dwyer, P. D., Gilkeson, J. H., and List, J. A., 2002. Gender differences in revealed risk taking: evidence from mutual fund investors. *Economics Letters*, *76*(2), pp. 151-158.

Eagly, A. H., and Carli, L. L., 1981. Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: a meta-analysis of social influence studies. *Psychological Bulletin*, 90(1).

Easterby-Smith, M. P., Thorpe, R., and Jackson, P., 2008. Management research: theory and research.

Eckel, C.C. and Grossman, P.J., 2008. Men, Women and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economics Results*, 1, pp. 1061-1073.

Fagley, N. S., and Miller, P. M., 1990. The effect of framing on choice interactions with risk-taking propensity, cognitive style, and sex. *Personality and Social Psychology Bulletin*, *16*(3), pp. 496-510.

Evans, N., 2012. A 'nudge' in the wrong direction. *IPA Review*, 64(4), pp. 17-19.

Federal Glass Ceiling Commission, 1995. *Solid Investments: Making Full Use of the Nation's Human Capital*. Washington, D.C.: U.S. Department of Labor.

Finucane, M. L., Slovic, P., Mertz, C. K., Flynn, J., and Satterfield, T. A. 2000. Gender, race, and perceived risk: The 'white male' effect. *Health, Risk & Society*, 2(2), pp. 159-172.

Fishburn, P. C., and Kochenberger, G. A., 1979. Two-Piece Von Neumann-Morgenstern Utility Functions*. *Decision Sciences*, *10*(4), pp. 503-518.

Fischhoff, B., 1983. Predicting frames. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), p. 103.

Ganzach, Y., and Karsahi, N., 1995. Message framing and buying behavior: A field experiment. *Journal of Business Research*, *32*(1), pp. 11-17.

Ghauri. P., and Grønhaug, K., 2010. *Research Methods in Business Studies*, fourth edition, Prentice Hall, Financial Times.

Gneezy, U., Leonard, K.L., and List, J.A., 2009. Gender differences in competition: evidence from a matrilineal and patriarchal society. *Econometrica,* 77(5), pp. 1637–1664.

Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P., 2004. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), p. 93.

Harbaugh, W. T., Krause, K., and Vesterlund, L., 2002. Risk attitudes of children and adults: Choices over small and large probability gains and losses. *Experimental Economics*, *5*(1), pp. 53-84.

Hartog, J., Ferrer-i-Carbonell, A., and Jonker, N. 2002. Linking measured risk aversion to individual characteristics. *Kyklos*, 55(1), pp. 3-26.

Hartman, R. S., Doane, M. J., and Woo, C. K., 1991. Consumer rationality and the status quo. *The Quarterly Journal of Economics*, pp. 141-162.

Haslam, S.A., and McGarty, C., 2004. *Experimental Design and Causality in Social Psychological Research*, in Carol Sansome, Carolyn C. Morf and A.T. Palmer (eds.), The SAGE Handbook of Methods in Social Psychology, Sage Publications, Inc. Thousand Oaks: California

Hasseldine, J., and Hite, P. A., 2003. Framing, gender and tax compliance. *Journal of Economic Psychology*, *24*(4), pp. 517-533.

Heijden, E.V., Klein, T. J., Müller, W., and Potters, J. J. J., 2011. *Nudges and impatience: Evidence from a large scale experiment.* Netspar Discussion Paper No. 09/2011-081.

Hershey, J. C., and Schoemaker, P. J., 1980. Risk-taking and problem context in the domain of losses: An expected utility analysis. *Journal of Risk and Insurance*, pp. 111-132.

Huang, Y., and Wang, L., 2010. Sex differences in framing effects across task domain. *Personality and Individual Differences*, *48*(5), pp. 649-653.

Jacobsen, B., Lee, J. B., Marquering, W., and Zhang, C. Y., 2014. Gender differences in optimism and asset allocation. *Journal of Economic Behavior & Organization*, 107, pp. 630-651.

Jianakoplos, N. A., and Bernasek, A., 1998. Are women more risk averse? *Economic Inquiry*, *36*(4), p. 620.

Johnson, J. E. V., and Powell, P. L., 1994. Decision Making, Risk and Gender: Are Managers Different? *British Journal of Management*, 5, pp. 123-138.

Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Häubl, G. and Weber, E. U., 2012. Beyond nudges: Tools of a choice architecture. *Marketing Letters*, 23(2), pp. 487-504.

Kahneman, D., Knetsch, J. L., and Thaler, R. H., 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98(6), pp. 1325-1348.

Kahneman, D., Knetsch, J. L., and Thaler, R. H, 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives*, pp. 193-206.

Kahneman, D. and Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica*, 47, pp. 263-291.

Kahneman, D., and Tversky, A., 1984. Choices, values, and frames. *American Psychologist*, *39*(4), p. 341.

Knetsch, J. L., and Sinden, J. A., 1984. Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *The Quarterly Journal of Economics*, pp. 507-521.

Knetsch, J. L., 1989. The endowment effect and evidence of nonreversible indifference curves. *The American Economic Review*, pp. 1277-1284.

Landsberger, H. A., 1958. Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry. *Distribution Center, N.Y.S. School of Industrial and Labor Relations, Cornell University, Ithaca, New York*

Levin, I. P., Schneider, S. L., and Gaeth, G. J., 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), pp. 149-188.

List, J. A., 2004. Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, *72*(2), pp. 615-625.

Litwin, M. S., 1995. How to measure survey reliability and validity (Vol. 7). *Sage Publications.*

Loewenstein, G., and Kahneman, D., 1991. Explaining the endowment effect. *Working paperDepartment of Social and Decision Sciences, Carnegie-Mellon University*.

MacCrimmon, K.R., and Wehrung, D.A., 1990. Characteristics of risk-taking executives. *Management Science*, 36, pp. 422-435.

Madrian, B. C., and Shea, D. F., 2000. *The power of suggestion: Inertia in 401 (k) participation and savings behavior* (No. w7682). National bureau of economic research.

Mason, W., and Suri, S., 2011. *Conducting behavioral research on Amazon`s Mechanical Turk.* Behav Res, 44, pp. 1-23.

Mechanical Turk Blog, 2011. *Mturk mail bag: How do I create HITs that require a Worker have an approval rate of 95% on MY HITs?* [Online] Available at: http://mechanicalturk.typepad.com/blog/2011/11/mturk-mail-bag-how-to-send-work-to-workers-with-high-approval-rates-on-my-hits.html [Accessed 3 Nov 2015].

Meyerowitz, B. E., and Chaiken, S., 1987. The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of Personality and Social Psychology*, *52*(3), p. 500.

Meyerson, P., & Tryon, W. W., 2003. Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, *35*(4), pp. 614-620.

Mook, D. G., 1983. In defense of external invalidity. *American Psychologist*, *38*(4), 379.

Mongin, P., 1997. Expected utility theory. Handbook of economic methodology, pp. 342-350.

Niederle, M., and Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, President and Fellows of Harvard College and the Massachusetts Institute of Technology.

Paolacci, G., 2012. Inside the Turk: Methodological Concerns and Solutions in Mechanical Turk Experimentation. *Advances in Consumer Research.* Volume 40.

Payne, J. W., Laughhunn, D. J., and Crum, R., 1980. Translation of gambles and aspiration level effects in risky choice behavior. *Management Science*, *26*(10), pp. 1039-1060.

Personvernombudet for forskning (NSD). *Opprett nytt meldeskjema*. [Online] Available at: http://www.nsd.uib.no/personvern/meldeskjema [Accessed 15 Sept 2015].

Policonomics, 2012. *Risk and uncertainty II: Risk aversion*. [Online] Available at: http://www.policonomics.com/lp-risk-and-uncertainty2-risk-aversion/ [Accessed 29 Sept 2015].

Pratt, J. W., 1964. Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society*, pp. 122-136.

Rademacher. J.D.M., and Lippke, S., 2007. Dynamic online surveys and experiments with free open-source software dynQuest. *Behaviour Research Methods*, 39 (3), pp. 415-426.

Ringdal, K., 2009. *Enhet og mangfold. Samfunnsvitenskapelig forskning og kvantitativ metode*. 2. utgave, Fagbokforlaget Vigmostad & Bjørke AS, Bergen.

Riva, G., Teruzzi, T., and Anolli, L., 2003. The use of the internet in psychological research: comparison of online and offline questionnaires. *Cyber Psychology & Behavior*, *6*

Ross, S. A., 1981. Some stronger measures of risk aversion in the small and the large with applications. *Econometrica: Journal of the Econometric Society*, pp. 621-638.

Samuelson, W., and Zeckhauser, R., 1988. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), pp. 7-59.

Saunders, M. N., Saunders, M., Lewis, P., and Thornhill, A., 2009. *Research methods for business students, 5/e*. Pearson Education.

Schmidt, G.B., 2015. Fifty Days an MTurk Worker: The Social and Motivational Context for Amazon Mechanical Turk Workers. *Industrial and Organizational Psychology*, 8(2), pp. 165-237.

Schmidt, U., and Traub, S., 2001. An experimental test of loss aversion. *Journal of Risk and Uncertainty*, *25*(3), pp. 233-249.

Schubert, R., Brown, M., Gysler, M. and Branchinger, H.W., 1999. Financial Decision-making: Are Women Really More Risk-Averse? *The American Economic Review*, 89(2), pp. 381-385.

Schurchkov, O., 2012. Under pressure: Gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association* 10(5), pp. 1189–1213, Wellesley College.

Shapira, Z., 1995. Risk taking: A managerial perspective. Russell Sage Foundation.

Slovic, P., Fischhoff, B., and Lichtenstein, S., 1982. Why study risk perception. *Risk Analysis*, *2*(2), pp. 83-93.

Sunden, A. E., and Surette, B. J., 1998. Gender differences in the allocation of assets in retirement savings plans. *American Economic Review*, pp. 207-211.

Sunstein, C. R., 2014. Nudging: a very short guide. *Journal of Consumer Policy*, 37(4), pp. 583-588.

Sunstein, C., and Thaler, R., 2008. *Nudge*. *Improving Decisions about Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.

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

Tversky, A., & Kahneman, D., 1991. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics,* 106(4), pp. 1039-1061.

U.S Census Bureau, 2013a. *Selected Social Characteristics in the United States.* [Online] Available at: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_5YR_DP02&src=pt [Accessed 12 Nov 2015].

U.S Census Bureau, 2013b. *ACS Demographic and Housing Estimates*. [Online] Available at: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_5YR_DP05&prodType=table [Accessed 12 Nov 2015].

U.S Census Bureau, 2014. *Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States.* [Online] Available at: http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=PEP_2014_PEPAGESEX&prodType=table [Accessed 12 Nov 2015].

Vinogradov, D., and Shadrina.E., 2013. Non-monetary incentives in online experiments. *Economics Letters,* 119, pp. 306-310.

Weber, E.U., Blais, A.R. and Betz, N.E., 2002. A Domain-specific Risk-attitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making*, 15, pp. 263-290.

Wooldridge, J.M., 2014. *Introduction to Econometrics, Europe, Middle East and African Edition*, Cengage Learning EMEA

Yates, J.F., and Stone, E.R., 1992. *The risk construct*. In Risk-taking Behavior, Wiley series in human performance and cognition, pp. 1-25. Oxford, England: John Wiley & Sons.

Zeffane, R., 2013. Gender, Trust and Risk-taking: A literature Review and proposed Research Model. *The Proceedings of the 9th European Conference on Management Leadership and Governance.* Klagenfurt: Academic Conferences and Publishing International Limited.

Appendix

A.1 Survey in Qualtrics

Introduction

Introduction
Welcome to this research project. We very much appreciate your participation.
This survey is conducted as a part of a master thesis project at the Norwegian School of Economics.

Participation
Participation in this research study is completely voluntary. You have the right to withdraw at any time or refuse to participate entirely without jeopardy to future participation in other studies conducted by us.

Confidentiality
All data obtained from you will be kept confidential and will only be reported in an aggregate format (by reporting only combined results and never reporting individual ones). All questionnaires will be concealed, and no one other than the primary investigators will have access to them.

Procedures
The study will take about 5-6 minutes to complete. Please always make sure to read the instructions carefully.

Payment
If you complete the survey, you will be entitled to your participation fee of 1 USD.

Your participation will be registered on your Amazon Mechanical Turk worker ID. You will not need a completion code. Your payment will be paid to you within two weeks after the completion of this HIT.

Questions about the research
If you have questions regarding this study, you may contact questions.preferencesurvey@gmail.com

I have read and understood the above consent form and desire to participate in this study.

- ○ Yes
- ○ No

Amazon Mechanical Turk worker ID

Please note that your participation will be registered on the following Amazon Mechanical Turk worker ID:

${e://Field/workerId}

The worker ID was retrieved automatically when you clicked on the link that brought you here. This step is necessary for assigning payments to the right account and to ensure that you only participate in this study once.

Instructions

Instructions

In the main part of the study, you will be working on a picture categorization task. You will see a picture on your screen and are asked to select the categories from the menu below the picture that you think fit to the picture and its content. You can select multiple categories from the menu if you think that the picture fits into that category as well.

Your answer will be submitted automatically after 30 seconds and you will be given a new picture. There are 6 pictures in total. The task will last for 3 minutes in total and we ask you to work thoroughly with each picture.

Click the button below if you have read and understood the instructions.

Picture categorization task

Picture 1: Please select the elements below that best describes the picture:

☐	People	☐	Jungle	☐	Warm
☐	Beach	☐	Boats	☐	Cold
☐	Forest	☐	Mountains	☐	Traveling
☐	Houses	☐	Night	☐	Sightseeing
☐	Cars	☐	Day	☐	Hiking
☐	Animals	☐	Sunny	☐	Shopping
☐	Plants	☐	Rainy	☐	Swimming
☐	Roads	☐	Cloudy	☐	Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds*

#QuestionText, TimingClickCount#: *0 clicks*

30

Picture 2: Please select the elements below that best describes the picture:

☐	People	☐	Jungle	☐	Warm
☐	Beach	☐	Boats	☐	Cold
☐	Forest	☐	Mountains	☐	Traveling
☐	Houses	☐	Night	☐	Sightseeing
☐	Cars	☐	Day	☐	Hiking
☐	Animals	☐	Sunny	☐	Shopping
☐	Plants	☐	Rainy	☐	Swimming
☐	Roads	☐	Cloudy	☐	Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds*

#QuestionText, TimingClickCount#: *0 clicks*

~~3 0~~

Picture 3: Please select the elements below that best describes the picture:

☐	People	☐	Jungle	☐	Warm
☐	Beach	☐	Boats	☐	Cold
☐	Forest	☐	Mountains	☐	Traveling
☐	Houses	☐	Night	☐	Sightseeing
☐	Cars	☐	Day	☐	Hiking
☐	Animals	☐	Sunny	☐	Shopping
☐	Plants	☐	Rainy	☐	Swimming
☐	Roads	☐	Cloudy	☐	Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds*

#QuestionText, TimingClickCount#: *0 clicks*

30

Picture 4: Please select the elements below that best describes the picture:

☐	People	☐	Jungle	☐	Warm
☐	Beach	☐	Boats	☐	Cold
☐	Forest	☐	Mountains	☐	Traveling
☐	Houses	☐	Night	☐	Sightseeing
☐	Cars	☐	Day	☐	Hiking
☐	Animals	☐	Sunny	☐	Shopping
☐	Plants	☐	Rainy	☐	Swimming
☐	Roads	☐	Cloudy	☐	Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds*

#QuestionText, TimingClickCount#: *0 clicks*

30

Picture 5: Please select the elements below that best describes the picture:

☐	People	☐	Jungle	☐	Warm
☐	Beach	☐	Boats	☐	Cold
☐	Forest	☐	Mountains	☐	Traveling
☐	Houses	☐	Night	☐	Sightseeing
☐	Cars	☐	Day	☐	Hiking
☐	Animals	☐	Sunny	☐	Shopping
☐	Plants	☐	Rainy	☐	Swimming
☐	Roads	☐	Cloudy	☐	Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds*

#QuestionText, TimingClickCount#: *0 clicks*

30

Picture 6: Please select the elements below that describes the picture:

☐ People	☐ Jungle	☐ Warm
☐ Beach	☐ Boats	☐ Cold
☐ Forest	☐ Mountains	☐ Traveling
☐ Houses	☐ Night	☐ Sightseeing
☐ Cars	☐ Day	☐ Hiking
☐ Animals	☐ Sunny	☐ Shopping
☐ Plants	☐ Rainy	☐ Swimming
☐ Roads	☐ Cloudy	☐ Eating

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

#QuestionText, TimingPageSubmit#: *0 seconds* #QuestionText,

TimingClickCount#: *0 clicks*

3 0

Payment

You have been working on the picture task for 3 minutes. As a reward for completing this work task, you are given a lottery ticket that gives you the chance to earn 250 bonus points or 0 bonus points with equal probability.

The bonus points are converted into **USD at a rate of 1 cent per bonus point**.

The bonus points that you have received by the end of this study will be paid to you using the bonus system within a few days after the completion of this HIT. Please click on the button to continue.

You have been working on the picture task for 3 minutes. As a reward for completing this work task, you were given a payment of 100 bonus points.

The bonus points are converted into **USD at a rate of 1 cent per bonus point**.

The bonus points that you have received by the end of this study will be paid to you using the bonus system within a few days after the completion of this HIT. Please click on the button to continue.

Possibility to change safe payment to lottery

Before we continue, you will have the possibility to decide whether you want to keep your payment of 100 bonus points or whether you want to change it for a lottery ticket. The lottery ticket gives you the chance to receive 250 bonus points or 0 bonus points with equal probability. Please indicate your choice below:

- ◯ Keep the safe payment
- ◯ Exchange the safe payment with a lottery

Possibility to change lottery to safe payment

Before we continue, you will have the possibility to decide whether you want to keep your lottery ticket and the chance to earn 250 bonus points or 0 bonus points with equal probability or whether you want to exchange it for a payment of 100 bonus points. Please indicate your choice below:

- ◯ Keep the lottery ticket
- ◯ Exchange the lottery ticket with a safe payment

Time preference

We now ask for your willingness to act in a certain way. Please indicate your answer on a scale from 1 to 7, where 1 means you are "completely unwilling to do so" and a 7 means you are "very willing to do so". How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4

- ◯ 5
- ◯ 6
- ◯ 7

Demographics

You have completed most of the survey. We would now like to ask you a few questions about your background before we conclude this survey.

What is your gender?

- ◯ Male
- ◯ Female

How old are you?

0 Years

In which of these regions do you usually live?

- ◯ Northeast (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania)
- ◯ Midwest (Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota)
- ◯ South (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington D.C., West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas)
- ◯ West (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington)
- ◯ Other

What kind of area do you usually live?

- ◯ Rural area
- ◯ Urban area

What is your ethnicity? If you belong to several groups, please indicate the one you identify with the most.

- ◯ White American
- ◯ European American
- ◯ Middle Eastern American
- ◯ Hispanic American or Latino American
- ◯ Black American or African American
- ◯ Native American or Alaske Native
- ◯ Asian American
- ◯ Foreign
- ◯ Other

What is the highest level of education you have completed?

- ○ Less than High School
- ○ High School / GED
- ○ Some College
- ○ 2-year College Degree
- ○ 4-year College Degree
- ○ Masters Degree
- ○ Doctoral Degree
- ○ Professional Degree (JD, MD)

Do you have any education within the field of economics?

- ○ No
- ○ Yes, I have taken one or a few courses
- ○ Yes, I have 1-2 years of education within economics or related fields
- ○ Yes, I have 3 or more years of education within economics or related fields

What is your current occupational situation?

- ○ Employed
- ○ Retired
- ○ Works in own household
- ○ Student
- ○ Freelance or self-employed
- ○ Unemployed
- ○ Other

What is your annual household income?

- ○ Below $10,000 USD
- ○ $10,000 - $20,000 USD
- ○ $20,000 - $30,000 USD
- ○ $30,000 - $40,000 USD
- ○ $40,000 - $50,000 USD
- ○ $50,000 - $60,000 USD
- ○ $60,000 - $70,000 USD
- ○ $70,000 - $80,000 USD
- ○ Above $80,000 USD
- ○ Do not know

What is your marital status?

- ○ Married
- ○ Unmarried partner
- ○ Divorced
- ○ Widowed
- ○ Single

How many children do you have?

- ○ No children
- ○ 1 child
- ○ 2 children
- ○ 3 children
- ○ 4 children
- ○ 5 or more children

Finally, if you have any comments or suggestions related to this study please write them down in the field below. Your feedback is important to our study.

End of survey

Thank you!

You have successfully finished the survey and we thank you for your participation!
We will calculate and pay your bonus as soon as this full batch of HITs is finished. It generally takes us up to two weeks to match the data and pay out the bonuses.

Double entries are a problem to us. Please do not try to take this survey again.
If you have any questions, rather contact us by sending an email to questions.preferencesurvey@gmail.com.

A.2 Documentation of experimental procedures

A.2.1 Control variables

The appropriate and necessary control variables was identified by reviewing existing literature. We investigated the control variables used in other equivalent studies and reviewed the criticism against the studies. Dohmen et al., (2011) find strong evidence that gender, age, height, and parental background play an important role in explaining individual differences in risk attitudes. In their meta-study from 2008, Eckel and Grossman argue that both field and lab studies typically fail to control for knowledge, wealth, marital status and other demographic factors that might bias the measures of gender differences in risk-taking behavior.

We want to control for knowledge, because there is evidence of the gender difference being (at least partly) explained by knowledge (e.g. Dwyer, Gilkeson and List, 2002). We include knowledge as a control variable by asking the participants about their level of education. We also ask whether they have any education within the field of economics. Furthermore, Sundén and Surette (1998) finds that marital status is significantly related to asset allocation, with married men and women being more risk averse than single men and women. For this reason, we found it interesting to include a question about marital status, to see whether this might have an impact. We also include a question about number of children, as this might influence risk-taking behavior (Jianakoplos and Bernasek, 1998). Moreover, we ask the participants about their ethnic background, as culture also is found to influence the gender gap (e.g. Finucane, Slovic, Mertz, Flynn and Satterfield, 2000).

A.2.2 Layout

Designing the experiment was an iterative process involving discussions and evaluations of everything from the layout and wording to technical details. As we only had one chance to carry out the experiment, due to budget constraints, it was crucial to make it right the first time. We reviewed an extensive amount of literature to find the best design of similar experiments. We also got access to some relevant prior studies conducted by The Choice Lab. This was the starting point for our discussions regarding the design. We developed several

outlines, and together with helpful guidance and insights from our supervisor and other researchers at The Choice Lab, we agreed upon the final design.

We focused on a convenient and tidy layout to facilitate the respondents' willingness to answer. In particular, we focused on the formulation of the instructions and questions to make sure the wording was clear, precise and simple. We tried to avoid terminology or concepts that could create confusion. In this regard, we evaluated the probable educational level, knowledge and cultural background of our intended sample to make sure the tasks and questions were understandable. This also ensured that we were effective in our communication, so that our intended meaning corresponded with the perceived meaning (Ghauri and Grønhaug, 2010).

A.2.3 Pre-test

Prior to the experiment, we examined the clarity and the procedures of the experimental design by talking to five friends and acquaintances about the design and wording of the experiment. This enabled us to be effective in producing the wanted information considering our time and budgetary constraints, compared to conducting an incentivized pre-test on a smaller sample (Ghauri and Grønhaug, 2010). The contributors conducted the experiment in Qualtrics. After completing the experiment, participants were given a printed transcript of the experiment and asked to write down their thoughts.

The participants were asked explicitly about their understanding of the questions and tasks, so that we could evaluate our wording. All the participants were able to complete the tasks and answer the questions, indicating that we are efficiently in our communication. The participants' own feedback revealed no confusion about the experiment. Even when asked explicitly about their interpretation of different parts, no confusion was uncovered. In addition to asking about clarity, we discussed potential ethical issues with the contributors, assuring that the experiment is justifiable from an ethical perspective. It can be argued that all people to a certain degree are ethical sources and by talking about issues regarding ethics with the contributors we can get some insight into how real participants will react.

Sensitivity was a central topic in our pre-test. By sensitivity, we refer to both sensitive information, such as social security numbers, and questions about sensitive topics. Examples

could be undesirable behavior like shop lifting or not caring about recycling. We asked the contributors whether they felt that the experiment and questions where sensitive, and whether or not they were easily willing to answer. All the contributors replied that they were not reluctant to answer any of the questions, and stated that they did not assess the experiment to be sensitive. This seems reasonable considering that the experiment does not ask about topics where participants might not want to reveal their true preferences. Facilitating willingness to answer is also addressed through the confidentiality assured in the first part of the experiment.

A third and final topic we asked the contributors to assess was the impact of the American culture on the questions asked. Here we focused the talk around the meaning and scope of the different concepts, ensuring that Americans understand and interpret the questions and tasks in the same manner. The contributors did not think that there would be any confusion here. The contributors should have some knowledge of the American culture considering that it is a part of the western world and that the world has become "smaller". Two of the contributors have lived in the U.S., and two others have traveled there. However, since none of the participants where Americans we cannot be sure that we have covered everything in regard to culture.

Two of the participants mentioned that they thought the Americans would be more reluctant to answer sensitive questions. In this regard, we discussed the questions about income, marital status and occupation, and whether we should include an option of "do not want to answer". We were reluctant to include such answer options as they give little value in our analysis. We concluded on not including this option as the participants has the possibility to end the study if they do not want to provide the requested information.

Furthermore, we discussed how the different backgrounds and levels of education might influence the understanding of the questions and the response accuracy. We reached the conclusion that the concepts are well explained by the use of structured questions, indirectly explained through the various response options. Hence, "the response format should explain the concepts so that they are correctly understood and interpreted and thereby encourage correct and clear answers" (Ghauri and Grønhaug, 2010, p. 222).

In addition, we measured the time it took to complete the survey to make sure we had a correct estimate of the duration. We were concerned that it would take too long to finish the

study compared to our estimated time and that it would demotivate respondents, and make them careless about their answers. It took approximately 5-7 minutes for all the contributors to conduct the study. This was as expected. We also asked the contributors specifically about the time restrictions on the picture task. All of the contributors thought it was enough time, and reporter that they did not feel rushed.

To conclude, since there were no discovered problems related to duration, understandability, sensitivity, format or culture we used the same layout and procedure later in the main study.

A.2.4 Technical specifications of the experiment

Before the experiment was executed on mTurk, it was ensured that the design of the treatment worked. More specifically, that all the participant was allocated evenly and randomly into one of the two treatment groups. This was made possible by features in Qualtrics that enabled us to create a "randomizer" to randomly allocate participants to one of two groups. In addition, one could chose to make the "randomizer" allocate an equal amount of individuals to each group, so that it would not be arbitrary how many got allocated to each group. This was important to us as an uneven distribution could affect our analysis and results.

To make sure that that the same individuals not conducted the survey more than once, we made use of a feature in Qualtrics called "Prevent ballot box stuffing". This feature places a cookie on the participant's browser when they submit a response. If they try to click on the survey link again, this cookie is detected and prohibit the individual from taking the survey. However, the participant can avoid this restriction by accessing the survey from different browsers. They may do this to try to get paid multiple times or to try to figure out what the survey was about. We try to deal with this problem by including the following statement at the end of our survey: "Double entries are a problem to us. Please do not try to take this survey again. If you have any questions, rather contact us by sending an email to questions.preferencesurvey@gmail.com". We did not experience any problems with double entries.

When the survey was closed, the responses in Qualtrics had to be matched with the workers on mTurk, to ensure that the participants got the participation fee (1 USD) they were entitled to. The participation fee is expected to be paid shortly after the completion of the HIT, so the

matching was done the same day. Due to confidentiality and security reasons, the payments were done by The Choice Lab, and not by us.

The participants could also earn a bonus in the experiment. To decide the bonus, we had to match each participant's worker ID and assignment ID with his or her response to the payment question. If the participant chose the safe payment, he or she would be entitled to a 1 USD bonus. If the participant chose lottery, we had to decide whether he or she won 2.5 USD or nothing. To make this decision fair and truly randomized, we made use of the "Rand()" formula in Excel. After listing the bonus to be paid to each participant, the transactions were again done by The Choice Lab.

A.2.5 Feedback from participants

During and after the experiment, we assessed the feedback from participants. This was provided to us through free text comment boxes both in the Qualtrics survey and on mTurk. No boredom or tiredness was reported by any of the participants. This was expected due to the short duration and monetary incentives provided. Most participants that made a comment, stated that everything was easy to read and understandable and that all the pictures loaded well and quickly. However, one participant reported a problem on one of the pictures. Another one reported that he/she wanted more time on the picture assignment. Two reported that it would have been easier to see and rate the pictures if we had put the categories to the right of the picture instead of below. However, this was not feasible as we wanted to make it possible to conduct the experiment on all devices, including tablets and phones. Furthermore, this should not have any consequences for the result of the study since the picture task has nothing to do with the real purpose of the study.

Another participant mentioned that we should have included an additional alternative for the question about living area. We only included the options "Rural" and "Urban". He/she suggested to also include "Suburb". This was a good suggestion and we agree that this would have been be smart to include. Regarding drop-out rates and participants taking the survey multiple times, we could not detect any problems.

A.2.6 Quality assurance

Mason and Suri (2011) states that the downside of using fast and cheap data is the potential of low quality. They claim that most workers on mTurk are not primarily motivated by financial returns. Still there are a few groups of workers who only care about the financial returns without considering their quality of work. These workers are characterized as *Spammers*. The attention level of workers on mTurk has been questioned and is a concern for those using mTurk to conduct surveys, experiments and research (Paolacci, 2012).

To identify Spammers in our study and test for the participants' attention level, we started by looking more closely at their answers on the picture categorization tasks (the work task). We identified six participants that had three or more none responses, out of the six picture tasks. One of these answered that he was zero years. Another participant had a very large variance in his answers, four blank picture tasks and 11 and 16 answers on the remaining tasks. In addition, he pick categories that was rarely stated by any of the other participants. We could not detect any irregularities among his background questions, although it is impossible to tell if a participant is being honest on these questions, as all answer options are reasonable.

Furthermore, we looked at irregularities within the remaining sample: If they had picked the most common categories or not, if they picked more than 15 categories on the same picture tasks, and if they had answered the control questions in a satisfying way. No further irregularities were detected and we thereby categorized the six participants who only answered two or less of the work tasks as spammers.

Even though we identified six participants to be careless on the picture task this does not mean that they were carless on the risk-taking task. This is reasonable to assume because they were incentivized on this task. This was verified by testing for the effects of spammers by running a t-test without spammers and compare with a t-test including the entire sample. There were no significant differences between the groups. The responses were therefore kept and included in the subsequent analysis in order to maximize the sample size. Consequently, we do not consider these spammers to be jeopardizing the validity of the results.

A.3 Ethical considerations

"It is the researcher's responsibility to ensure that appropriate steps are taken to conduct ethical research" (Mason and Suri, 2011, p. 15). Research should follow some ethical guidelines and make sure that respondents are treated in an ethical manner. To ensure ethical research we report our findings, methods and instruments accurately and honestly, so that the readers can make judgments about the reliability of our findings (Ghauri and Grønhaug, 2010). In the following sections, we discuss issues related to online experiments in general, although our focus will be on issues particularly relevant to the use of mTurk. As the use of mTurk is a more resent phenomenon, the related ethical issues are still under debate (Mason and Suri, 2011). We want to highlight informed consent, debriefing, restricted populations, compensations and confidentiality as central ethical considerations.

A.3.1 Informed consent

To make sure the participants know what they are participating in, we have an information page that explains different aspects of the experiment, with a consent question at the bottom of the page. This consent allows participation without a written signature (Bryant, Hunton and Stone, 2004). Here the subjects are explained the purpose and scope of the experiment. We tried to be as honest as possible without revealing too much of the purpose of the experiment. This was done deliberately to avoid participation biases. We therefore decided not to inform the subject about the main purpose to measure risk-taking behavior. We still consider the provided scope and purpose information to be adequate enough for the subjects to make an informed decision.

The information page further describes participation rights, that participation is completely voluntarily and that participants can withdraw at any time. It also includes information about estimated duration, confidentiality and benefits in the sense of payment. An email address is provided so that the subjects can contact us if they have any questions or comments. With the provision of this information subjects should be able to make an informed judgment about whether they want to participate or not (Mason and Suri, 2011; "American Psychological Association", 2015; Ghauri and Grønhaug, 2010)

A.3.2 Debriefing

At the end of the survey, participants are asked to write down comments or suggestions if they have any questions or feedback. The use of the comment box provided insight into the participants' reflections regarding the study, helping us to uncover problems in the design, and potential high drop-out rates. No further information or debriefing about the purpose of the research is stated. We do not consider debriefing to be an important issue regarding our research because there are no deceptions (including brand or product crises) or undisclosed information in the study (Mason and Suri, 2011). Hence, we do not consider our survey to provide any harm, distress or confusion on the participants ("American Psychological Association", 2015; Ghauri and Grønhaug, 2010).

A.3.3 Restricted populations

Mechanical Turks policy states that "use of the Amazon Mechanical Turk site is limited to persons and entities that lawfully can enter into and form contracts under applicable law. It is not intended for use by minors" ("Mechanical Turk", 2015). This policy is enforced by requiring that the payments are linked to verifiable U.S account, making it harder for minors to be accepted as workers. Unfortunately, this cannot be entirely avoided because minors can use other people's identity to be accepted as a worker (Mason and Suri, 2011).

A.3.4 Compensation

Ethical consideration regarding compensation on Mechanical Turk is debated due to the low wages the workers receive. Workers on Mechanical Turk fall outside the minimum wage laws because they are considered "independent contractors". In defense of mTurk, it could be argued that the participation is completely voluntarily, and that the working hours and working conditions are decided by the participants themselves (Mason and Suri, 2011). Mason and Suri (2011) also states that most workers do not rely on payment from this platform as a necessity, which suggest that low payments can be justifiable.

Another argument posed by Mason and Suri (2011, p.16) is that workers on Mechanical Turk are self-selected making a "marked for lemons", and thereby argue that "the equilibrium wage is lower than if the requester could more easily check the quality of work before

compensating the workers". Considering this, we decided to reward the participants no lower than the standard payment used on the platform.

A.3.5 Confidentiality

Workers on Mechanical Turk are provided a worker ID that does not contain personally identifiable information. However, there are some issues regarding the storage of data obtained on Mechanical Turk (Mason and Suri, 2011). With the use of this platform, Amazon will have access to the data. We avoid this issue by using an external HIT, where the responses are stored in Qualtrics rather than on mTurk. The participants in our study are ensured confidentiality through the reporting of the result only in an aggregated format (by reporting only combined results and never reporting individual ones). Nobody else than the primary investigators and those helping us conducting the experiment has access to the data.

By using the worker ID, provided by Amazon, and an assignment ID, which is necessary to pay the participants, the information stored cannot be traced back to individuals. The built-in payment on mTurk enables us to preserve anonymity at the same time as we provided adequate incentives (Duersch et al., 2009; Mason and Suri, 2011). The data gathered does not contain name, personal number or any other personal characteristics. The data can neither be traced back to an e-mail or an IP-address. In our Qualtrics survey we made sure to activate the "anonymize responses" function, so that it would not be possible to track the responses back to the participants.

It is not possible for us to link the participants' answers to their identity through the descriptive data because of the general nature of these questions. In summary, we conclude that we exclusively register anonymous information and satisfy confidentiality policies outlined by "Personvernombudet" [8] (Personvernombudet for forskning, 2015). We will only use the data for the purpose that we have stated and we will be careful regarding the storage of data. From the workers perspective, a guarantee of anonymous treatment of responses will be reassuring and may influence the participant to take the survey seriously. This will hopefully encourage participants to provide genuine data, which is crucial for our research (Ghauri and Grønhaug, 2010).

[8] «Personvernombudet» is part of the Norwegian Social Science Data Services, and is responsible for monitoring the privacy policies on behalf of the research industry of Norway (Personvernombudet, 2015).

A.4 Descriptive statistics

Gender

#	Answer	Response	%
1	Male	199	55%
2	Female	161	45%
	Total	**360**	**100%**

Time preference

#	Answer	Response	%
1	1	3	1%
2	2	6	2%
3	3	18	5%
4	4	65	18%
5	5	121	34%
6	6	83	23%
7	7	64	18%
	Total	**360**	**100%**

Geography*

#	Answer	Response	%
1	Northeast	60	17%
2	Midwest	93	26%
3	South	148	41%
4	West	58	16%
5	Other	1	0%
	Total	**360**	**100%**

Urban

#	Answer	Response	%
1	Rural area	138	38%
2	Urban area	222	62%
	Total	**360**	**100%**

Ethnicity

#	Answer	Response	%
1	White American	287	80%
2	European American	13	4%
3	Middle Eastern American	0	0%
4	Hispanic or Latin American	13	4%
5	Black Am. or African Am.	20	6%
6	Native Am. or Alaske Native	4	1%
7	Asian American	21	6%
8	Foreign	0	0%
9	Other	2	1%
	Total	**360**	**100%**

Education

#	Answer	Response	%
1	Less than High School	3	1%
2	High School / GED	54	15%
3	Some College	96	27%
4	2-year College Degree	42	12%
5	4-year College Degree	135	38%
6	Masters Degree	27	8%
7	Doctoral Degree	1	0%
8	Professional Degree (JD, MD)	2	1%
	Total	**360**	**100%**

Eco_education

#	Answer	Response	%
1	No	220	61%
2	Yes, I have taken one or a few courses	113	31%
3	Yes, I have 1-2 years of education within economics or related fields	14	4%
4	Yes, I have 3 or more years of education within economics or related fields	13	4%
	Total	**360**	**100%**

Occupation

#	Answer	Response	%
1	Employed	222	62%
2	Retired	5	1%
3	Works in own household	19	5%
4	Student	12	3%
5	Freelance or self-employed	73	20%
6	Unemployed	26	7%
7	Other	3	1%
	Total	**360**	**100%**

Income

#	Answer	Response	%
1	Below $10,000 USD	19	5%
2	$10,000 - $20,000 USD	47	13%
3	$20,000 - $30,000 USD	66	18%
4	$30,000 - $40,000 USD	54	15%
5	$40,000 - $50,000 USD	40	11%
6	$50,000 - $60,000 USD	32	9%
7	$60,000 - $70,000 USD	31	9%
8	$70,000 - $80,000 USD	20	6%
9	Above $80,000 USD	50	14%
10	Do not know	1	0%
	Total	**360**	**100%**

Marital status

#	Answer		Response	%
1	Married		143	40%
2	Unmarried partner		52	14%
3	Divorced		21	6%
4	Widowed		1	0%
5	Single		143	40%
	Total		**360**	**100%**

Children

#	Answer		Response	%
1	No children		213	59%
2	1 child		55	15%
3	2 children		48	13%
4	3 children		29	8%
5	4 children		10	3%
6	5 or more children		5	1%
	Total		**360**	**100%**

*Geography: The four regions are explained below with all states included in each region.

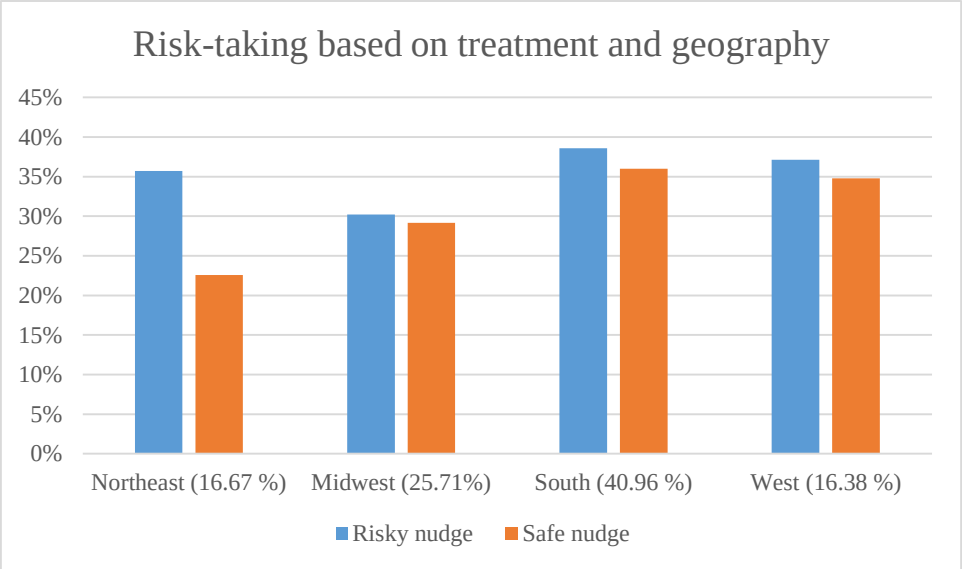
Northeast: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania

Midwest: Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota

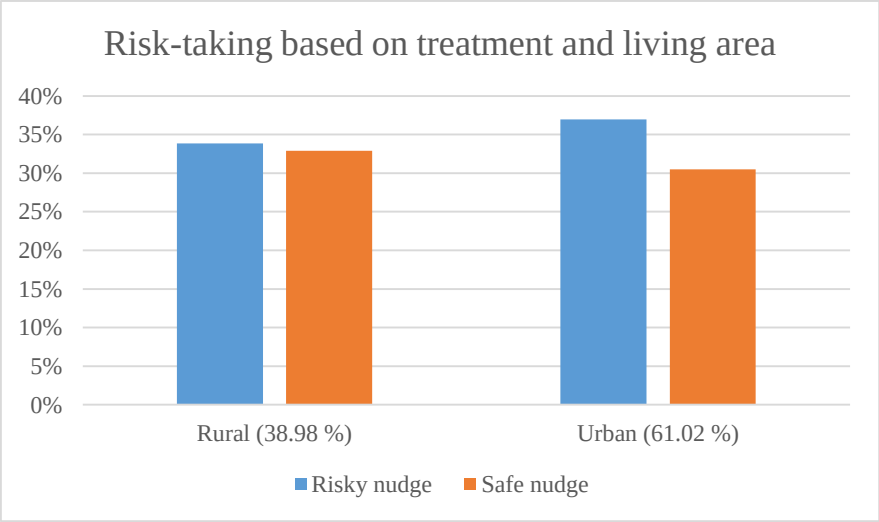
South: Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, Washington D.C., West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas

West: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington

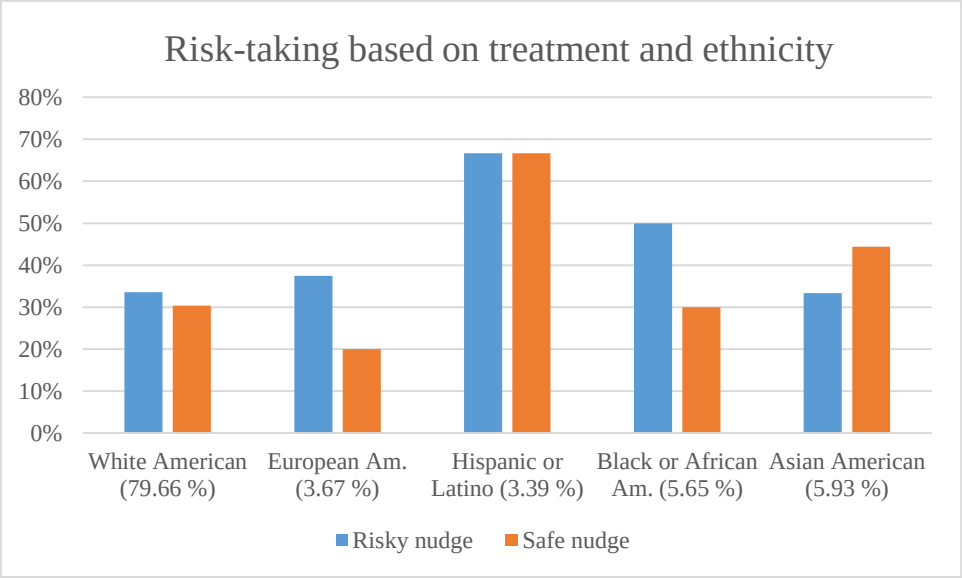
A.5 Risk-taking based on treatment and background variables



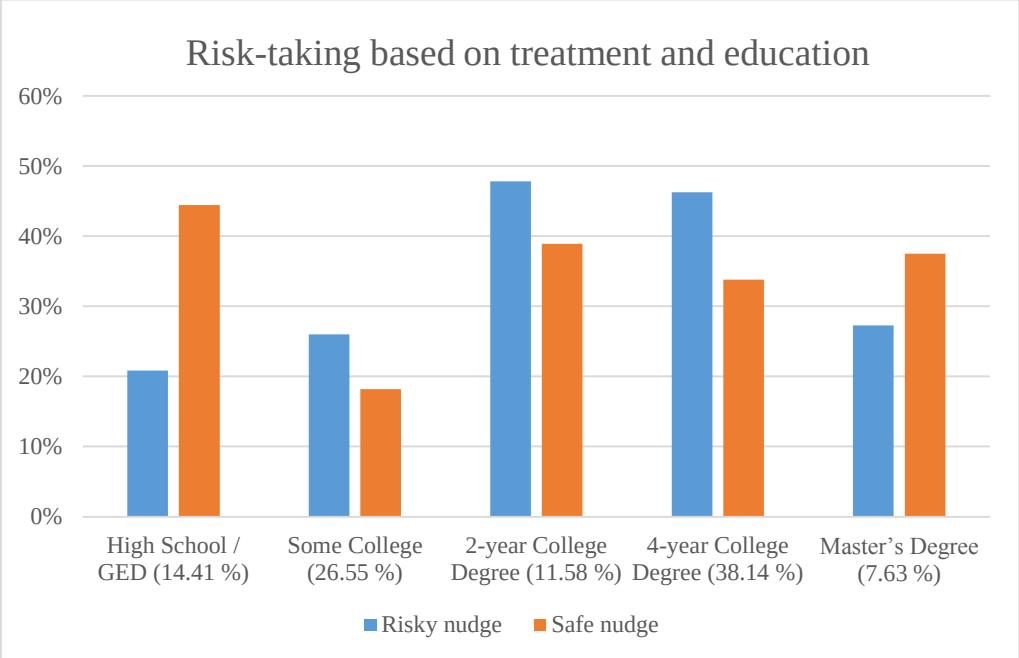
Geography	Risky default	Safe default
Northeast (16.67 %)	35.71 % risk seeking (n: 28)	22.58 % risk seeking (n: 31)
Midwest (25.71%)	30.23 % risk seeking (n: 43)	29.17 % risk seeking (n: 48)
South (40.96 %)	38.57 % risk seeking (n: 70)	36.00 % risk seeking (n: 75)
West (16.38 %)	37.14 % risk seeking (n: 35)	34.78 % risk seeking (n: 23)
Other (0.28 %)	- (n: 0)	0.00 % risk seeking (n: 1)



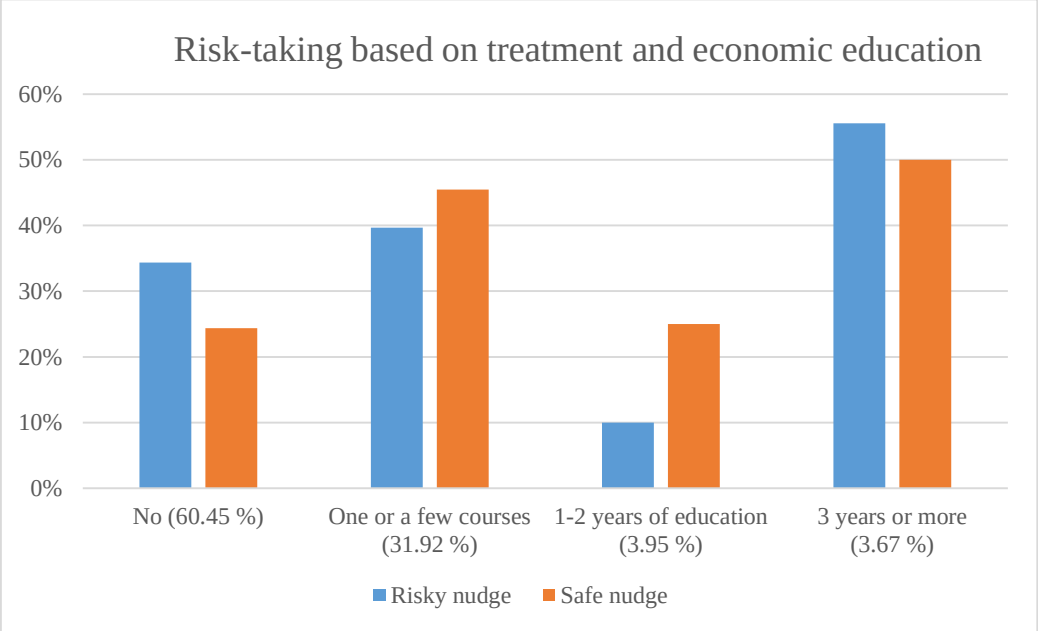
Living area	Risky default	Safe default
Rural (38.98 %)	33.85 % risk seeking (n: 2)	32.88 % risk seeking (n: 1)
Urban (61.02 %)	36.94 % risk seeking (n: 3)	30.48 % risk seeking (n: 3)



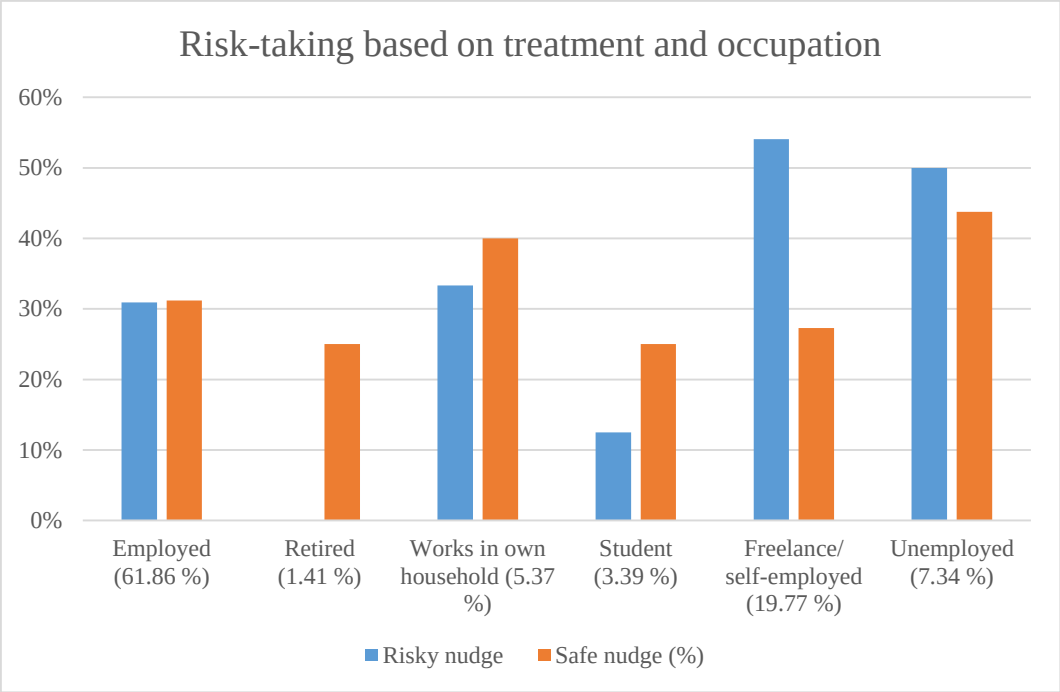
Ethnicity	Risky default	Safe default
White American (79.66 %)	33.58 % risk seeking (n: 137)	30.34 % risk seeking (n: 145)
European American (3.67 %)	37.50 % risk seeking (n: 8)	20.00 % risk seeking (n: 5)
Hispanic or Latino (3.39 %)	66.67 % risk seeking (n: 6)	66.67 % risk seeking (n: 6)
Black or African Am. (5.65 %)	50.00 % risk seeking (n: 10)	30.00 % risk seeking (n: 10)
Native Am. or Alaska (1.13 %)	0.00 % risk seeking (n: 2)	0.00 % risk seeking (n: 2)
Asian American (5.93 %)	33.33 % risk seeking (n: 12)	44.44 % risk seeking (n: 9)
Other (0.56 %)	100.00 % risk seeking (n: 1)	0.00 % risk seeking (n: 1)



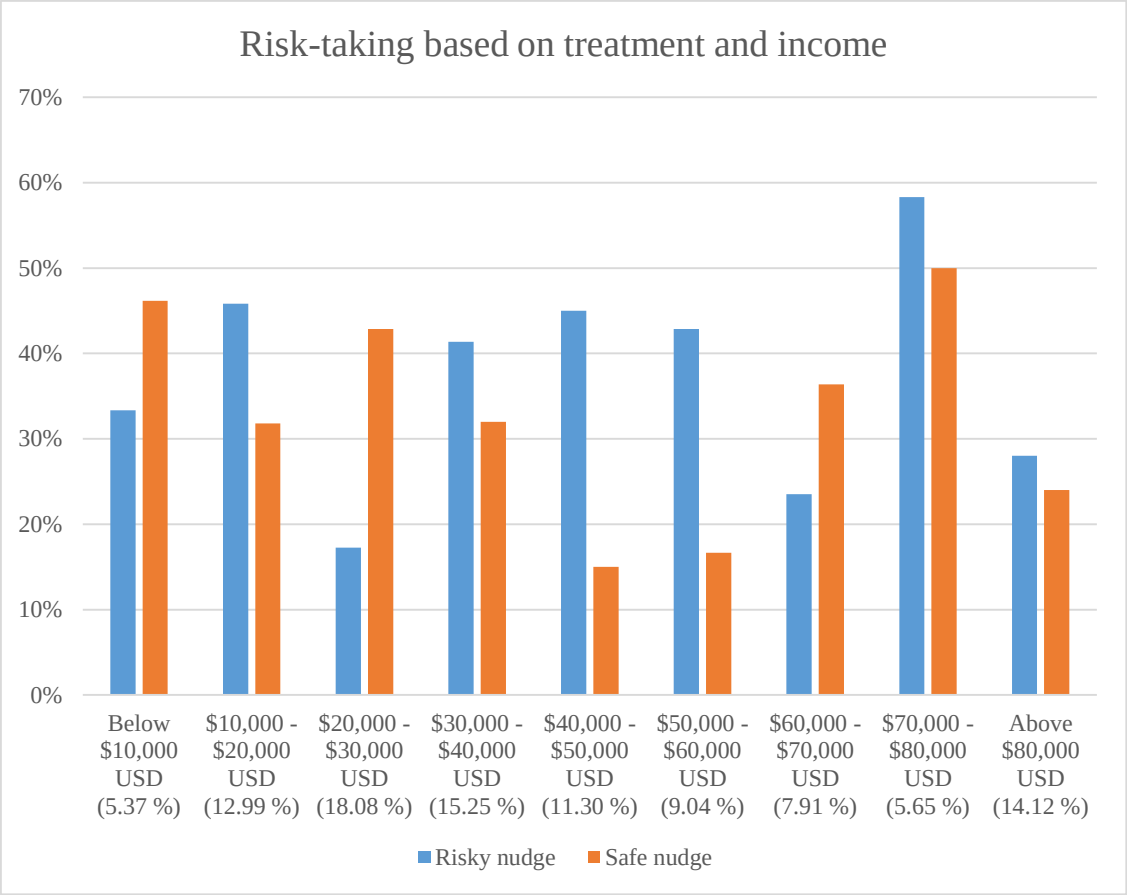
Education	Risky default	Safe default
Less than High School (0.85 %)	- (n: 0)	0.00 % risk seeking (n: 3)
High School / GED (14.41 %)	20.83 % risk seeking (n: 24)	44.44 % risk seeking (n: 27)
Some College (26.55 %)	26.00 % risk seeking (n: 50)	18.18 % risk seeking (n: 44)
2-year College Degree (11.58 %)	47.83 % risk seeking (n: 23)	38.89 % risk seeking (n: 18)
4-year College Degree (38.14 %)	46.27 % risk seeking (n: 67)	33.82 % risk seeking (n: 68)
Master's Degree (7.63 %)	27.27 % risk seeking (n: 11)	37.50 % risk seeking (n: 16)
Doctoral Degree (0.28 %)	- (n: 0)	0.00 % risk seeking (n: 1)
Professional Degree (0.56 %)	0.00 % risk seeking (n:1)	0.00 % risk seeking (n: 1)



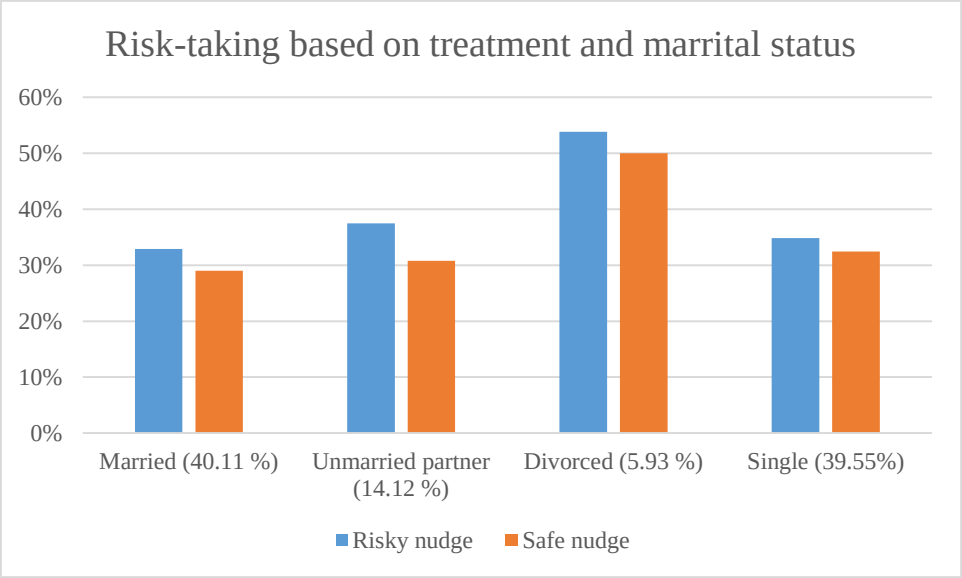
Economic education	Risky default	Safe default
No (60.45 %)	34.34 % risk seeking (n: 99)	24.35 % risk seeking (n: 115)
One or a few courses (31.92 %)	39.66 % risk seeking (n: 58)	45.45 % risk seeking (n: 55)
1-2 years of education (3.95 %)	10.00 % risk seeking (n: 10)	25.00 % risk seeking (n: 4)
3 years or more (3.67 %)	55.56 % risk seeking (n: 9)	50.00 % risk seeking (n: 4)



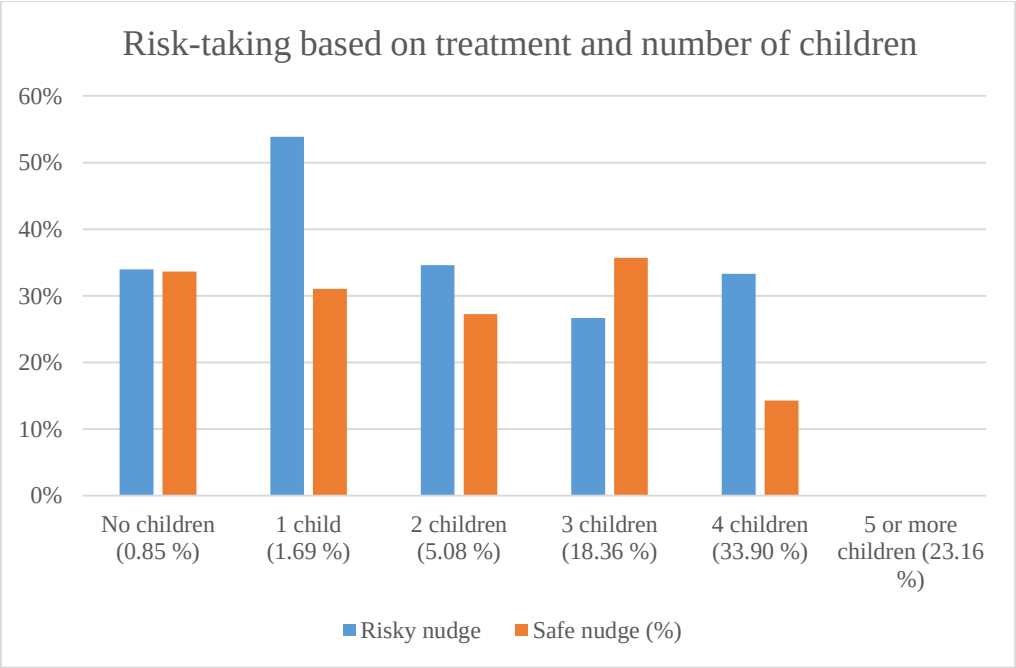
Occupation	Risky default	Safe default
Employed (61.86 %)	30.91 % risk seeking (n: 110)	31.19 % risk seeking (n: 109)
Retired (1.41 %)	0.00 % risk seeking (n: 1)	25.00 % risk seeking (n: 4)
Works in own household (5.37 %)	33.33 % risk seeking (n: 9)	40.00 % risk seeking (n: 10)
Student (3.39 %)	12.50 % risk seeking (n: 8)	25.00 % risk seeking (n: 4)
Freelance /self-employed (19.77 %)	54.05 % risk seeking (n: 37)	27.27 % risk seeking (n: 33)
Unemployed (7.34 %)	50.00 % risk seeking (n: 10)	43.75 % risk seeking (n: 16)
Other (0.85 %)	0.00 % risk seeking (n: 1)	0.00 % risk seeking (n: 2)



Income	Risky default	Safe default
Below $10,000 USD (5.37 %)	33.33 % risk seeking (n: 6)	46.15 % risk seeking (n: 13)
$10,000 - $20,000 USD (12.99 %)	45.83 % risk seeking (n: 24)	31.82 % risk seeking (n: 22)
$20,000 - $30,000 USD (18.08 %)	17.24 % risk seeking (n: 29)	42.86 % risk seeking (n: 35)
$30,000 - $40,000 USD (15.25 %)	41.38 % risk seeking (n: 29)	32.00 % risk seeking (n: 25)
$40,000 - $50,000 USD (11.30 %)	45.00 % risk seeking (n: 20)	15.00 % risk seeking (n: 20)
$50,000 - $60,000 USD (9.04 %)	42.86 % risk seeking (n: 14)	16.67 % risk seeking (n: 18)
$60,000 - $70,000 USD (7.91 %)	23.53 % risk seeking (n: 17)	36.36 % risk seeking (n: 11)
$70,000 - $80,000 USD (5.65 %)	58.33 % risk seeking (n: 12)	50.00 % risk seeking (n: 8)
Above $80,000 USD (14.12 %)	28.00 % risk seeking (n: 25)	24.00 % risk seeking (n: 25)
Do not know (0.28 %)	- (n: 0)	0.00 % risk seeking (n: 1)



Marital status	Risky default	Safe default
Married (40.11 %)	32.88 % risk seeking (n: 73)	28.99 % risk seeking (n: 69)
Unmarried partner (14.12 %)	37.50 % risk seeking (n: 24)	30.77 % risk seeking (n: 26)
Divorced (5.93 %)	53.85 % risk seeking (n: 13)	50.00 % risk seeking (n: 8)
Widowed (0.28 %)	- (n: 0)	0.00 % risk seeking (n: 1)
Single (39.55%)	34.85 % risk seeking (n: 66)	32.43 % risk seeking (n: 74)



Children	Risky default	Safe default
No children (0.85 %)	33.98 % risk seeking (n: 103)	33.65 % risk seeking (n: 104)
1 child (1.69 %)	53.85 % risk seeking (n: 26)	31.03 % risk seeking (n: 29)
2 children (5.08 %)	34.62 % risk seeking (n: 26)	27.27 % risk seeking (n: 22)
3 children (18.36 %)	26.67 % risk seeking (n: 15)	35.71 % risk seeking (n: 14)
4 children (33.90 %)	33.33 % risk seeking (n: 3)	14.29 % risk seeking (n: 7)
5 or more children (23.16 %)	0.00 % risk seeking (n: 3)	0.00 % risk seeking (n: 2)

A.6 Validity and reliability

A valid and reliable research design is important to enhance the trustworthiness of our results and reduce the possibility of errors in our research (Saunders et al., 2009). In this section, we assess how reliable and valid our research is, in order to evaluate the quality of our study. The objective is to minimize the amount of error so that our data provide a more accurate reflection of the truth (Litwin, 1995). Simultaneously, we will evaluate the strengths and weaknesses of our study. We will first discuss and evaluate four categories of validity, namely internal-, external- construct- and statistical conclusion validity. Thereafter, we will assess the reliability of our experiment.

A.6.1 Validity

A.6.1.1 Internal Validity

Internal validity is concerned with the causal relationship between the dependent and independent variables (Ghauri and Grønhaug, 2010). Thus, internal validity is concerned with whether or not the results imply what they are intended to predict (Ringdal, 2009). As our data is based on an online experiment, as opposed to laboratory testing, there may be variables influencing the dependent variable (risk taking) that is not in our control. In this regard, it has been argued that the environmental characteristics are more variable (Saunders, et al., 2009). This includes the technical aspects of the equipment, noise and lighting (Dandurand et al., 2008).

Online experiments may also be more prone to adverse effects of distractions. It is likely that participants get sidetracked more easily or have been working on other tasks simultaneously while conducting an experiment online. This can contribute to decreased accuracy (Dandurand et al., 2008; Mason and Suri, 2011). However, as this may be true for more extensive studies that have higher demands for concentration, cognitive capacity and attention, this is likely not to be a substantial threat in a simple, short and incentivized experiment (Dandurand et al., 2008).

Another possible threat that might influence the quality of the results is the prevalence of *spammers* on mTurk (see Appendix A.2.6 about Quality assurance) (Mason and Suri, 2011).

These workers are, however, probably not careless on the risk-taking task, due to the fact that this task is incentivized. Hence, we do not consider this to be a problem on the obtained results. In addition to this, it can be argued that mTurk is less inclined to have spammers in their subject pool, than other crowdsourcing sites, due to a built-in reputation system for workers (Mason and Suri, 2011).

A substantial threat to internal validity is the possibility of multiple submissions (Dandurand et al., 2008). With the use of an external HIT, the threat of multiple submissions is problematic because participants only need to change their browser to submit a new survey (see section A.2.4). We tried to prevent this by including a request not to participate more than once. This might have mitigated the problem although we cannot be sure that it has eliminated it (Bryant et al., 2004; Mason and Suri, 2011).

It could also be specific events (history) occurring prior to or during the experiment, or high drop-out of the experiment affecting the results (Ghauri and Grønhaug, 2010; Bryant et al., 2004; Dandurand et al., 2008; Mason and Suri, 2011). Because the data is cross-sectional, i.e. the data is gathered in a certain point in time (1-2 hours), history effects are not a major threat. Experimental mortality becomes a threat to internal validity when there is something special about participants dropping out of the experiment as compared to those that complete the study. This could therefore create a participant self-selection bias that rivals the explanation for the observed finding. (Bryant et al., 2004). Even though we are incapable of detecting drop-outs, we do not consider drop-out rates to be a large problem in our study. This is mainly due to the fact that our survey is incentivized and that the participants do not receive payment unless they complete the survey (Mason and Suri, 2011). Providing contact information for questions and pre-testing of instructions may also have reduced drop-outs (Mason and Suri, 2011).

An online experiment may be less inclined to diffusion and imitation of treatments. The higher probability of a more geographically disbursed population should make it harder for one participant in the treatment group to learn information intended only for those in the other treatment group (diffusion). Online experiment should also be less inclined to have participants in one treatment condition imitating those in the other treatment condition (imitation). Facilitating this is the build in control of hiding the browser back button, which prevents participants to go back and change their answers. Including the picture assignment as

"warm-up" task should also reduce the adverse effects of drop-out due to the fact that drop-outs should occur before the random assignment to conditions (Dandurand et al., 2008; Bryant et al., 2004).

On the other hand, the targeted population on mTurk seems to be interacting online through the use of communities, where they share information and opinions with each other (Schmidt, 2015; Paolacci, 2012). It is also possible that workers learn from their nearest network of friends about this platform, making interaction among mTurk workers more likely. Paolacci (2012), however, finds this to not be a critical issue on mTurk and conclude that cross-talk hardly can contribute substantially to participant non-naivety. In this regard, it is also unlikely that participants have had time to communicate due to the fact the study was just accessible for around two hours. In relation to this, it is important to note that participants recruited could have conducted experiments that are conceptually or methodologically related to our experiment. With the possibility of conduction an unlimited number of experiments, this could have a negative impact on quality or accuracy of our survey results (Paolacci, 2012; Schmidt, 2015).

By the use of two treatment groups and by randomization we can investigate the relative effects. We should thereby be able to remove the possibility of systematic differences between the groups, preventing third factors distorting the effect of gender on risk-taking (i.e. the only different between the two groups are the treatments). This is further strengthened by testing for control variables. The direction and causality between gender and risk is also verified by the extensive literature search. With this, threats to internal validity should be minimized and we should be well suited to draw a causal conclusion from the obtained correlation among variables. Thus we consider the internal validity of our research to be strong (Saunders et al., 2009).

A.6.1.2 External Validity

External validity refers to whether the findings can be generalized beyond the particular study at hand, to other contexts, populations or periods in time (Ghauri and Grønhaug, 2010; Saunders et al., 2009). The use of an online experiment strengthens the external validity because it is carried out in a more natural decision making environment compared to a laboratory experiment (Vinogradov and Shadrina, 2013). Comparative advantages in this

sense are no pressure of an artificial laboratory environment, taking the experiment whenever is convenient, less time pressure and possibly greater work life balance, imposing less stress and more comfort on participants (Bryant et al., 2004). Thus, online experiments can be done in a wider array of contexts, not just in the highly concentrated context of the laboratory (Dandurand et al., 2008).

Using mTurk as a sampling frame further enables us to select from a larger, more heterogeneous population than we otherwise would be able to reach (Bryant et al., 2004; Duersch et al., 2009). It allows us to have a more diverse population with varying age and socioeconomic status, living in different geographical regions. This broadens the sample beyond the standard subject pools (Rademacher and Lippke, 2007; Dandurand et al., 2008). The use of mTurk as the sampling frame should therefore increase the generalizability of the results compared to our alternative which was undergraduate students at our university.

To be able to draw general conclusions, it is critical that the sample is representative of the population it is supposed to predict an effect on. Whether our findings can be generalized to other countries that may differ in terms of resources, labor conditions, culture and traditions is difficult to predict. It is more likely that the findings can be generalized to other western cultures. More importantly, we want to assess whether the results can generalize to the American population. The fact that we use a non-probability, self-selection sampling technique and mTurk as a sampling frame might make generalizability more difficult.

Furthermore, participants who previously have conducted experiments from The Choice Lab can choose to get a notice when The Choice Lab post new surveys, making us prone to an even higher selection bias. Another source of selection bias can be drop-outs, as mentioned in the section about internal validity. As elaborated upon in section 3.2.4 about our sample, our sample is younger, more male dominated, higher educated and includes less ethnic minorities than the general US population. Thus the sample is rather similar to the American population, although not perfectly representative. We can thereby not be sure that participants systematically differ from non-participants.

If our findings can be generalized to other situations or periods in time is difficult to say. The experimental setting is often criticized to be artificial or unrealistic, advocating low applicability to other situations (e.g. Mook, 1983). The time between treatment and

measurement could also have an impact. The amount of time passing from one is exposed to the default to the choice regarding risk-taking is made is minimal. It is not certain how the default effect will unfold over time. Studies investigating such cases might find other results. However, compared to a traditional sampling we consider our external validity to be satisfying and at least as applicable to the American population. This can also be reasonably inferred from the demographics profile of the obtained sample.

A.6.1.3 Construct Validity

Construct validity addresses the concern of establishing the correct operational measures for the concepts being studied. (Ghauri and Grønhaug, 2010). Ghauri and Grønhaug (2010) highlight three characteristics or sources to construct validity. With our use of only one indicator for each concept, neither convergent nor discriminant validity can be assessed. However, we asses face validity to be strong in our study considering that we have consulted literature and our supervisor, assuring that the measure used seems to be reasonable for what we intend to measure (Ghauri and Grønhaug, 2010).

Misinterpretation of concepts and terms used in the study is further minimized due to thoroughly assessing the wording and instruction used. Clarity was also stressed by focusing on this in the pre-test, and verified by the obtained feedback from the participants (see section A.2.5). Another possible threat to validity is the test effect (Ghauri and Grønhaug, 2010). Sometimes, the experiment itself and the fact that the workers answers are being reported have an effect on their provided answers. This is called the Hawthorne effect (Landsberger, 1958). The Hawthorne effect also addresses the issue of people tending to alter their answers to how they think the researchers want them to answer. Nevertheless, this effect should be reduced because we explicitly assure the participants that everything is confidential. The "naturalism" of online experiments may also increase construct validity by decreasing demand effects and other experimenter influences. In this regard, it has been argued that participants online are less prone to altering their answers due to not meeting with experimenter(s) personally (Dandurand et al. 2008). Furthermore, the pre-testing of the experiment helps to avoid leading and charged questions.

A.6.1.4 Statistical Conclusion Validity

Statistical conclusion validity is the extent to which conclusions drawn about effects or causal relations is reflecting a true effect in the population or whether they are simply due to random events (Bryant et al., 2004). To prove statistical conclusion validity we assess our study's statistical power, significant testing and effect size.

As mentioned in the subchapter about power calculations, statistical power is a function of sample size, population effect size and α error. An increase in statistical conclusion validity is possible through the availability of a larger sample size (Mason and Suri, 2011). With the use of internet data collection we get a substantially larger sample size than we otherwise would be able to obtain. This increases our study's statistical power in comparison to our best possible alternative.

The larger sample size also lowers the likely error in generalizing to the larger population. The sample is more likely to be representative of the population from which they are drawn and, in particular, "the mean calculated for the sample is more likely to equal the mean for the population" (Saunders et al., 2009, p. 218). On the other hand, our obtained sample size is not optimal taking power calculations into consideration. The distribution of the sample among men and women and the two treatments further narrows the sample. As this sample is relatively small, the risk of committing a Type II[9] error increases. In turn this might lead to non-significant results. Thus, if the sample where larger we would probably have obtained results with higher significance due to the higher possibility of detecting small effects in data.

Another threat to statistical power is the distinction between men and women. This could result in an uneven sample size, where one of the groups has a significant larger sample than the other. In our sample there is some unevenness in the groups. This unevenness is not optimal but should not be large enough to be significant. However, with a small and uneven sample size, even when assigning the participants randomly, there is a higher possibility that the detected effects can be attributed to differences in the composition of the two groups.

[9] "Conclude that something is not true, when in reality it is" by accepting the null hypothesis (Saunders, et al. 2009, p. 452)

Furthermore, if the sample size is too small, possible outliers would have too much influence on the data, resulting in spurious results (Wooldridge, 2014). This occurs if the population from which the sample is drawn is not normally distributed. However, since our population in each gender category is above 30, the sampling distribution for the mean should be close to a normal distribution (Saunders et al., 2009).

In an online experiment the more "natural" or heterogeneous experimental setting may decrease statistical conclusion validity by increasing random error. When the sample is relatively heterogonous, as opposed to the classic student sample, it is more prone to variance caused by uncontrolled factors. As discussed above this "natural" setting can have a positive impact on the construct- and external validity. Hence, there is a tradeoff between a larger sample obtained through the use of online experiments and increased noise from lessened control over data collection. The question here is if the larger sample size reduces beta error sufficiently to compensate for the increase in noise.

An increase in statistical conclusion validity is obtained through minimizing random human data entry and transcription errors. As opposed to paper-and-pencil experiments, our study prevents participants from entering invalid responses by a build in control. The data is immediately downloaded to STATA which minimize transcript errors. Furthermore, the build in control for restricted time, drop-outs and forced answering prevents challenges with incomplete experiments that possibly could have rendered the obtained sample size (Bryant et al., 2004).

When it comes to significance testing, it is important to select the appropriate statistical tests for testing significance as well as specifying an appropriate significance level. The significant level can be defined as specifying the acceptable level of risk for rejecting the null hypothesis when it is in fact true (Type I error) (Saunders et al., 2009). We have used t-tests and linear multiple regressions as statistical tests. These are widely used and considered to be appropriate for significance testing.

Regarding the significance level we only found results that where significant at the 10 % level. This is a less stringent significant level than the widely used and accepted level of 5 %. This will increase the risk that we have committed type I errors, but decrease the risk that we have committed a type II errors. A factor that may violate statistical conclusion validity is

measurement error. As mentioned in the section about construct validity, we have used measurements that already have been applied in previous research. This should therefore not threaten the statistical conclusion validity of our study. We consider the conclusion validity in our results to be acceptable.

A.6.2 Reliability

Reliability refers to the stability of the results, and relates to the consistency of the research (Ghauri & Grønhaug, 2010). If the research is conducted again with the same measurement instruments and under the same conditions, the same results should be revealed. This implies that the study is reliable when random errors are removed (Ringdal, 2009).

A distinction is made between internal reliability and external reliability. Since we only use one item to measure one construct we are only going to address the issue of external reliability (Bryman and Cramer, 2009). External reliability of a study refers to the consistency of the measures over time (Bryman and Cramer, 2009). If the same respondents will respond the same to the same measurements at a later point in time, the external reliability is high. According to Saunders et al., (2009) there are four threats to external reliability; participant error, participant bias, observer error and observer bias.

Firstly, when it comes to observer bias we have minimized this by using closed questions which avoids subjective interpretation. In addition to this, we also avoid possible errors due to manually plotting of the data as the data were directly imported from Qualtrics into STATA. STATA also helps us to conduct automatic calculations. Further, the use of the software Qualtrics and conducting the experiment online facilitated uniformity of the study among participants (Dandurand et al., 2008). By including a forced response function in Qualtrics we also prevent respondents from submitting incomplete responses. This high level of structure should minimize observer errors.

We tried to eliminate participant error by choosing a "neutral" time of the day and the week. Participants on mTurk tend to be working most between Tuesday and Saturday. The time of the day when most workers are active is between 6 am and 3 pm (Mason and Suri, 2011). Our study was conducted in the middle of this period, namely on a Thursday. To make the survey available around noon in the US, we posted the survey in the afternoon in Norway. This was

also done to assure that the study was accessible at a convenient time and that a substantial part of the American population was awake, which also impacts the generalization of the results. We further tried to avoid participant errors by formulating the questions and instructions in the experiment carefully.

Participant bias may occur if respondents adapt their answers according to what they believe is the "correct" answer or what they believe the researcher is looking for. We tried to limit this by securing anonymity and by not directly informing the participants about the objective of the study. This issue was discussed more thoroughly in section about construct validity.

Based on the arguments and discussions in this subchapter, we evaluate the reliability of our study to be good. We have reported the exact procedure of our research as this provides the opportunity to replicate the research to verify the results (see Appendix A.2 for documentation). Hence, the measures used should be reliable to use for future research.