Overconfidence or Something Else:

Are People Genuinely Overconfident?

Name: Marcel Sbrizaj

Student number: 431513ms

MSc Economics and Business: Behavioural Economics specialisation

Supervisor: Dr. Chen Li

Date: 15.08.2016

Abstract

The question whether overconfidence is truly a psychological bias has attracted considerable attention in the overconfidence literature. Several studies have suggested that people only appear overconfident as a result of the experimental design, or some biases other than overconfidence. This paper researched the existence of genuine overconfidence. The term refers to a psychological bias that makes people believe they are better than average, overestimate their skills and knowledge and be excessively confident in their beliefs. However, several other biases could also produce the same effect and make people *appear* overconfident. This thesis investigated those biases. Then, it tried to measure overconfidence while minimizing their impact. Consequently, a new approach to measure overconfidence was needed. The new approach presented in this thesis is composed of two parts. In the first part, a typical general knowledge test is replaced by a cognitive game. The aim of the carefully designed cognitive game is to improve the often criticised standard method of measuring overconfidence and to mitigate the effects of other biases as much as possible. In the second part, incentivized choices are used to elicit subjects' beliefs. Measuring beliefs with incentivized choices gives participants a financial incentive to state their true beliefs and incentivized choices are rarely used in overconfidence experiments. Results from the experiment show that even when given monetary incentives, participants were not more successful in placing their performance relative to others. Moreover, subjects were overconfident in all overconfidence measures. In addition to overplacing their performance they also overestimated it and provided overly narrow confidence intervals, which support several of the genuine overconfidence hypotheses. In spite of the findings, the limitations of this study prevent us from comfortably making the conclusion that overconfidence is indeed a psychological bias.

Contents

Tables

"Confidence is good but overconfidence always sinks the ship."
Oscar Wilde

1. Introduction

Let's start by answering a simple question. What is overconfidence? The most popular definition is that overconfidence is a behavioural bias that makes people overestimate their abilities, skills and knowledge, underestimate risk and volatility of potential outcomes, and be excessively confident in their beliefs. Researchers that study overconfidence frequently assume that the different manifestations of overconfidence are a product of the same psychological bias (Moore and Healy 2008). However, sometimes we observe situations where people are overconfident and underconfident at the same time. How can we explain this? Soll (1996) pointed out that overconfidence is not necessarily attributed to psychological bias but could in fact be a product of the experimental environment. Other researches have also questioned overconfidence from a psychological standpoint.

The purpose of this paper was to study *genuine overconfidence* or overconfidence that is the result of a psychological bias. Genuine overconfidence is difficult to observe as there are biases other than overconfidence that could also produce the same effect. As a result, I introduced a new approach to measure overconfidence in order to study genuine overconfidence in isolation from those biases. This new approach is composed of two parts. In the first part, a general knowledge test, used by most overconfidence studies, was replaced by a cognitive game. The purpose of the cognitive game is to control for other factors as much as possible. In the second part, the exchangeability method of Baillon (2008) with incentivized choices is used to elicit subjects' beliefs. Incentivized choices are rarely used in overconfidence experiments.

First part of the thesis explains what overconfidence is and identifies all possible sources that could cause subjects to appear overconfident. Second part of the thesis describes the methodology and a different approach to measure overconfidence. The cognitive game and the exchangeability method, an elicitation technique to measure subjective probabilities, are described there. Results section presents the results of the experiment. The interpretations and the implication of the results are reviewed in the discussion section. Finally, the conclusion completes the thesis.

2. Literature Review

Overconfidence is a behavioural bias that makes people overestimate their abilities, skills and knowledge, underestimate risk and volatility of potential outcomes, and be excessively confident in their beliefs. To my knowledge, one of the earliest reports of overconfidence is the 1950 Brier Score developed by Glenn W. Brier. He tested the accuracy of weather predictions by comparing probability statements to actual occurrence of weather events. He tested for what is now known as miscalibration, one of overconfidence manifestations. Currently in the literature, overconfidence has three main manifestations.

1.) *Overestimation*. This measure of overconfidence refers to one's judgement about his or her[1] performance or skills. When a person overestimates his performance it means that he believes his performance was better than it actually was.
2.) *Overplacement*. This measure of overconfidence refers to one's evaluation about his performance relative to others. Overplacement is often called *better-than-average* effect and it means that a person believes he is better than the others.
3.) *Overprecision*. This measure of overconfidence refers to excessive confidence in one's beliefs. It occurs when people provide very narrow confidence intervals for their answers. In the literature, overprecision is often referred to as *miscalibration*.

In this paper I will try to search for *genuine overconfidence*. With this term I refer to a psychological element that makes people believe they are better than average and that they possess better skills and attributes.

The question whether overconfidence is a product of a psychological bias has received considerable attention in recent years. In fact, several researches have questioned overconfidence from a psychological standpoint. For instance, Soll (1996) pointed out that overconfidence is not necessarily attributed to psychological bias but could in fact be a product of the experimental environment. He insinuated that people only appear overconfident because experimenters create an environment that is unrepresentative of the real one. Larrick et al. (2007) had similar reservations about overconfidence being a psychological bias. They found that varying the difficulty of the task produces an opposite effect on different overconfidence measures. They advocated that those measures should be related, not going in the opposite

[1] From here on I will be using the male form throughout the entire paper. For example, sentences using the word *his* are actually referring to both men and women, but to achieve better flow, I will be using just the male form henceforth.

directions. Moore et al. (2008) reported that overconfidence studies have yielded inconclusive results. They pointed out that conclusions from those studies are ambiguous, because overconfidence has been studied in inconsistent ways. Correspondingly, Klayman et al. (1999) concurred that standard methods to measure overconfidence are not very useful to differentiate between different overconfidence manifestations. The main problem with interpreting overconfidence is that identifying its sources is extremely difficult (Keren 1997). Many different sources could make subjects appear overconfident but whether subjects are genuinely overconfident is up to debate. Limited attention has been devoted to explaining these sources.

In understanding overconfidence we must first ask ourselves whether different manifestations of overconfidence discussed above, can be attributed to different sources. Literature is not entirely clear on this. For instance, Moore et al. (2008) and Klayman (1999) suggested that different manifestations of overconfidence might have a different underlying cause. However, most researchers assume that different manifestations of overconfidence are related to the same psychological bias. To understand this argument we must first see how these overconfidence manifestations emerge and how they are connected.

Can a genuinely overconfident person be overconfident using one measure and underconfident using the other? In other words, can a person who thinks he is better than average be underestimating his performance at the same time? It has been observed that different overconfidence manifestations sometime go in opposite directions. But is that possible with genuine overconfidence? Let’s explore this phenomenon using the following example.

Imagine a person who believes he is an above average driver. But now also imagine, that this person underestimates his driving abilities. How can this be possible? Is this person underconfident or overconfident? One explanation is that he is underconfident. If he is a really good driver, then believing you are better than average is not overconfidence if you are actually better than average. This person could still be underplacing his performance, even though he believes he is better than average, if for example he believes he is only better than 60% of other drivers but is in reality better than 80% of other drivers. In that case this person would be underconfident in both measures. But what if that person is really better than only 40% of other drivers which is an indication of overplacement? It might be that this person is using his own definition of a good driver (discussed in detail on page 9-11), and thus correctly believes he is an above average driver (by his own definition). If overconfidence is psychological in nature, then I do not believe that different measures of overconfidence could go in opposite direction. I believe that a genuine overconfident person will believe he is an above average driver and also believe his driving abilities are better than they actually are. This is why identifying all the possible sources of overconfidence is so important and is the key to my research of studying genuine overconfidence. The next section discusses all the possible sources that could cause subjects to appear overconfident.

2.1 Sources of overconfidence

2.1.1 Miscalibration

According to Ben-David et al. (2013) miscalibration is defined as "excessive confidence about having accurate information" (p.2) or excessive precision in one's beliefs. Miscalibrated subjects provide confidence intervals that are too narrow and subjects frequently underestimate volatility of potential outcomes.

Miscalibration has been widely documented and is considered as evidence for overconfidence, but whether miscalibration is a result of the psychological factor is not very clear. Psychological bias could definitely be one of the reasons. If a person is genuinely overconfident, he will overestimate his skills and be excessively confident in his judgements. That will result in him providing narrow confidence intervals, which we observe as miscalibration.

But is that the only reason why a subject could appear miscalibrated? There are several biases that could cause this same effect. Indeed, if subjects were affected by those biases, then it is not necessarily the case, that they are genuinely overconfident. For instance, Kahneman and Tversky (1973) showed that subjects are consistently incapable of correctly evaluating frequencies and probabilities. This has nothing to do with a psychological bias that makes people believe their abilities are better than they actually are.

In a standard miscalibration experiment participants are normally given general knowledge questions. They are then asked to estimate the probability that their answers are correct, or in other words, to provide their confidence intervals. If their stated confidence intervals i.e. the proportion of answers they think they answered correctly, is higher than the actual proportion of correct answers, then they are miscalibrated. However, several contradictory findings appear in the literature. The difficulty effect or the hard-easy paradox is one of them. This was documented by Soll (1996) and Brenner et al. (1996) and they show that miscalibration is observed with hard tasks but the opposite-underconfidence is observed with easy tasks. Furthermore, they show that results are dependent on the way overconfidence is elicited. If subjects truly believe their abilities are better, then this should not be observed. If they are genuinely overconfident then changes in the research method should not influence their self-perception. This suggests that overconfidence might be more a product of the experimental environment.

Furthermore, Stankov and Crawford (1997) documented that the structure of the task can influence the correlation between expressed confidence and accuracy scores. In their study, they used a vocabulary test and a line lengths perceptual task, to measure overconfidence. The measured correlation between accuracy and confidence ratings for vocabulary scores was significantly different and higher than correlation obtained from the line lengths test. Stankov and Crawford (1997) explained that "People who are better at knowing the meanings of different words tend to be more confident about their knowledge" (p. 104). This bias is known as *illusion of knowledge*. It states that confidence increases at a faster pace than the accuracy, which makes subjects appear miscalibrated. The illusion of knowledge is not present in the line length test because there is no information about accuracy for subjects to process, unless subjects had some kind of prior experience in solving that type of exercise. In the vocabulary test, the more the subjects believed they knew about the meaning of words, the more likely they were to overestimate their score on the test. This means that any method using questions of knowledge is susceptible to this bias because it is impossible to know in which areas a particular subject is knowledgeable at. Overconfidence obtained from such a test is ambiguous because we cannot determine whether subjects genuinely believed their abilities are better, or it was the illusion of knowledge that made it appear so. Illusion of knowledge is a situational bias whereas genuine overconfidence is imbedded in a person's character and personality.

Incorrectly evaluating frequencies and likelihoods of events is another reason why subjects could appear miscalibrated. Kahneman and Tversky (1973), offered several explanations for how this could happen. A very common bias is representativeness, which they defined as "judging whether an event or sample is representative based on similarity of sample to the parent population" (Kahneman and Tversky, 1973). Two biases are generated by representativeness. Sample size neglect, and base rate neglect. Kahneman and Tversky reported that large classes and frequent events are recalled better. They suggested that the ease with which instances of an events are brought to mind is affecting how we estimate the probability of those events happening. If previous events are recalled quickly, we will most likely overestimate the probability of them happening again in the future. Suppose that we ask a New Orleans resident to estimate the probability of a new hurricane hitting the city in the next 10 years. The resident could instantly remember hurricane Katrina and may estimate the probability of another hurricane hitting the city to be quite high. Imagine now, that the weather is kind to the city of New Orleans and nothing happens in the next 10 years. What can we say about overconfidence and the New Orleans resident? I believe it is much more likely that the resident only appeared overconfident because he misjudged the probability of another hurricane happening due to representativeness and availability bias, as opposed to him being genuinely overconfident.

Subjects could also appear miscalibrated if they lie. Perhaps a bold statement but subjects might lie unintentionally. Keren (1997) documented that social norms is highly valued in today's society. He adds that social norms encourage overconfidence and that it is not surprising that we observe this phenomenon in many different areas, especially in standard overconfidence experiments with general knowledge questions. Participant might feel pressured or they do not want to appear unknowledgeable and thus lie about their confidence intervals.

2.1.2 Better-than-average effect

According to Benoît, Dubra & Moore (2015), *better-than-average* effect or *overplacement* is defined as an "effect in which a strict majority of people claim to be more adept in some domain than half the population" (p. 294). This effect has been reported in many different studies across multiple domains such as cognitive abilities, social skills and even looks and driving abilities.

Perhaps the most famous study in the field of overconfidence is Svenson (1981) *(Are we all less risky and more skilful than our fellow drivers?)* where he found that 92.7% of people believed they were above average concerning driving skills. Moreover, 46.3% of subjects rated themselves in the top 20%. At a first glance, it looks as a clear case of irrationality and overconfidence. However, this is not necessarily the case. As I will show in this section, in certain situations it is completely rational to rate yourself above average.

Furthermore, several contradictory findings appear in the better-than-average literature which makes interpretation of such literature somewhat limited. Particularly, Kruger & Dunning (1999) found that the above average effect is most pronounced in easy tasks, however the effect is actually reversed for hard tasks, where people believe they are below average. This is completely opposite of the hard-easy paradox in miscalibration studies, where people were more miscalibrated in hard tasks. Larrick, Burson and Soll (2007) argued that people have general views about themselves which they apply across different domains. They claimed that an overconfident person will be overconfident regardless of the task and domain. It follows, that the observed hard easy paradox and the reversed better-than-average effect cannot be attributed to genuine overconfidence alone and that several biases must play a role in creating the observed effect. I strongly agree with their argument and in fact, my definition of genuine overconfidence is based on that line of reasoning. A psychological bias that I associate with genuine overconfidence cannot choose where to appear. If people have a tendency to inflate their self-perception, it does not really make sense that a person would inflate their self-perception when evaluating his own performance but would go in the complete opposite direction when comparing that performance to the others. Therefore, I believe that a genuinely overconfident person is overconfident regardless of which manifestation of overconfidence we are talking about.

Roy & Liersch (2013) found that people have a "different, idiosyncratic definition for what it means to be a "good" driver" (p. 1649). Because there is no universal definition of a good driver, each person constructs his own perception of the skills a good driver should have. For one person, a definition of a good driver could be someone who always drives below the speed limit, is never involved in any accidents and drives so to minimize any risk associated with the road. For someone else, a good driver could be someone who is able to drive at 200km/h, able to drift and use a handbrake while driving around the curves. Using your own definition of a good driver, it is not so irrational that people believe they possess above average driving skills. Roy & Liersch (2013) and Sedikides & Gregg (2008) documented that the better-than-average effect is more pronounced when a task or a skill in question is more ambiguous.

People have their own idiosyncratic definition of a good driver but they also acknowledge that their definition of a good driver is different than the definition of someone else. What is more is that they acknowledge that under the different definition, they would not score as high as under their own (Roy and Liersch, 2013). This suggests that people may actually be quite capable of correctly assessing their skills, but they appear overconfident because they construct their own views and definitions that fit their best skills and abilities. Imagine now, we were to ask the same group of people from the Svenson experiment to evaluate their driving abilities, only this time to rank themselves based on how they would perform on a race track. Overconfidence would most likely be diminished, as the definition of what constitutes a good driver and a specific driving skill in question is clear (being able to drive fast). If subjects indeed use their own idiosyncratic definitions and are otherwise able to correctly assess their performance, then they are not genuinely overconfident.

Aside from idiosyncratic definition, there are other possible explanations for the "above average driver" effect. Evans (1991) reported that the majority actually drive regularly without getting into accidents. Because of availability heuristic and sample size neglect, a subject could overestimate the probability of an average driver to be involved in an accident. If the subject has not been involved in an accident (which is very likely), he might very well rate himself as an above average driver, based simply on the misjudged likelihood of one being in an accident. Additionally, the misanthropy effect postulates that a bias in memory can create negative perception of others (Ybarra, 1999, p. 261). This can further enhance the perception that others are bad drivers.

Better-than-average effect in other domains should also be questioned. Benoît et al. (2015) discovered that overplacement is not irrational if subjects formed their beliefs in a Bayesian manner. They explained that it might be completely rational for even the entire sample to bet on them being above average.

Another cause for the better-than-average effect could also be self-enhancement. It is hard to distinguish between self enhancement and overconfidence because they both produce exactly the same effect. I distinguished between the two by identifying self-enhancement as a rational utility generating process and genuine overconfidence as an irrational outcome of the psychological bias. If subjects act as utility maximizing agents, then choices done under self-enhancement are not irrational, even though they might appear to be on the outside. I will support this clam with the following studies that have identified self-enhancement as a utility generating activity.

People are generally motivated to possess a positive self-image because it is satisfying to hold one (Hepper, Gramzow, and Sedikides 2010 and Baumeister et al. 2003). Ybarra (1999) found a connection between utility and self enhancement and documented that the need to self-enhance can be removed. Furthermore, Sedikides and Gregg (2008) reported that self-enhancing happens only in domains that matter to people and that self-enhancing could be tactical. Moreover, they showed that people evaluate their work less positively, if they know they will have to justify their evaluation. If people know their evaluations are upward biased, then this could explain why self-enhancement is diminished in the presence of accountability. Indeed, Tice et al. (1995) documented that self enhancement is reduced among friends and is more pronounced among strangers, supporting the claim that accountability diminishes self-enhancement and that people actually know the true self. All this implies that people are not necessarily genuinely overconfident. They self-enhance because it makes them feel good and is giving them utility. The fact that self enhancement can be eliminated also suggests that it is not a psychological bias. A psychological bias cannot be removed so easily, as it is imbedded in a person's character and personality.

Furthermore, self-enhancement is not universal. Better-than-average effect is only observed in western countries, whereas in eastern countries this effect is almost non-existent (Lee, 2012). This author argues that "Asian cultures are markedly focused on self-improvement" (p. 270). Self enhancement would distort the true information about the self, making it harder to self-improve. In that context, self-enhancement does not create any utility and is therefore not present. In contrast, Americans place big emphasize on self-regard and can get more utility from self-enhancing (Lee, 2012, p.263 and 270).

Bénabou and Tirole (2005) documented that "self-enhancement might contribute to better success because it makes people exert more effort." Moreover, Creswell et al. (2007) reported that self-enhancement helps coping with stress and that it has genuine health benefits, further pointing to the fact that self-enhancement does create positive utility. When designing the experiment, I hypothesized that incentivizing choices would eliminate the utility to self-enhance. I based this hypothesis on the assumption that subjects prefer money to self-enhancing.

3. Methodology

A typical overconfidence experiment consists of subjects answering multiple choice questions, usually with two possible answers, related to general knowledge. Then, for each question, subjects assign a probability that their given answer is correct. Keren (1997) pointed out that those tests are long and boring and that it is unreasonable to assume subjects stay focused throughout the entire procedure. Additionally, he called the procedure highly artificial and unrepresentative of the real environment. He claimed that the interpretation of such data is ambiguous and generalizations made from those standard experiments are limited at best. Moreover, such a test is susceptible to many different biases which preclude us from observing genuine overconfidence. One could argue that the procedure I proposed could be subjected to the same criticism. It is indeed hard to design an experiment that would perfectly mimic the real environment and the experiment I proposed does not really score better in that department. However, it does score better in other departments. It is shorter and more interesting, or so I would like to think at least. But most importantly, its aim is to control for other factors so that we might be better equipped to observe genuine overconfidence. How that is achieved is discussed in this section.

As already mentioned, this paper presents a new approach to measure overconfidence. Specifically, I am talking about measuring genuine overconfidence. As such, it is not really important which manifestation do we want to measure, as they do not differ under the definition of genuine overconfidence. They are all a product of the same psychological bias. However, all manifestations can actually be tested with this design and all three manifestations (overprecision, overplacement and overestimation) were in fact tested in the experiment. The new designed consists of two parts. In the first part, a general knowledge test, used by most overconfidence studies, is replaced by a cognitive game. The second part concerns the elicitation methods to measure overconfidence. Different methods were used, most important of which is the exchangeability method (Baillon 2008) with incentivized choices, which are seldom used in overconfidence experiments. The purpose of the new design is to control for other biases as much as possible. If the game is successful in reducing the effect of other factors, then this new approach may be better suited to study genuine overconfidence.

3.1 Part 1 - The Cognitive Game

Experiment was designed and distributed using the Qualtrics software. The cognitive game consisted of 9 exercises. Exercises were chosen based on the pre-testing results. Determining factors were time needed to solve the exercise and the difficulty of the exercise. All the exercises were taken from the free public domain websites (28, 34, 35, 37, 48 in the references). Subjects were told that all participants would be ranked based on their total time. Time needed to solve the exercise was recorded using Qualtrics timing feature. The shorter the time, the better the rank. Subjects were also told that an extra 2 minutes would be added to their total time for each incorrect answer. There are several reasons why precisely 2 minutes was chosen. First, it allowed me to gather more exhaustive data, as people unable to solve the exercises were still able to finish the game (if they did not know the answer they guessed) and rank themselves later on. Second, while 2 minutes is a disadvantage, it is not as big a disadvantage that it would influence subjects' overconfidence levels. If the penalty was 10 minutes for example, we would probably not observe any overconfidence with people receiving those penalties. With 2 minutes however, I postulated that the penalty would not influence subjects' confidence levels. Moreover, on average subjects in the pretesting needed less than 2 minutes to solve the exercises and most of the subject gave up on an exercise after 2 minutes, if they were unable to solve it. Penalty of 2 minutes was chosen so that subjects guessing the exercises would not rank better than subjects actually solving them.

The design of the experiment serves several purposes. One of them is time. Standard methods to elicit overconfidence are long and boring and subjects' short attention spend is one of the main problems associated with those methods (Keren 1997). With standard general knowledge questions, usually at least 30 questions have to be asked in order for the overconfidence measure to be satisfactory. By using time as the central variable, the process does not need to be this long. It gives the experimenter more freedom and flexibility to design the experiment. Furthermore, several overconfidence measures can be tested at the same time. Finally the game was considered by many participants to be quite fun, which is an additional advantage of this approach.

The channel used to distribute the game was social media and participants played the game on their computers. As already mentioned, Keren (1997) pointed out that it is unreasonable to assume subjects stay focused throughout the entire procedure of the test. He was referring to the subject taking the general knowledge test in a controlled environment. My subjects were playing the game on their computers at home and it is safe to assume that their attention spend in such a setting is even smaller. For that reason I did not want the game to last more than 10 minutes and I chose the exercises, so that the combined time needed to solve them would not on average exceed that limit. I pretested and recorded the time for about 20 exercises, out of which 9 were selected for the game using this 10 minutes criteria.

I paid special attention to subjects exerting effort. I considered it extremely important for several reasons. Namely, if subjects are deliberately not giving any effort, if they are not using their skills, then there is no need for them to be overconfident about those skills. To make sure subjects exert effort, I chose exercises with different levels of difficulty. Some exercises were quite hard. Furthermore, I rewarded good performance in the game. Subjects were told that one participant playing the game would be selected at random, out of the pool of lotteries, and paid out €30. They were also told that the better their performance, the better their chances of winning the lottery. Each subject received 1 lottery for playing the game but could obtain additional lotteries based on their performance in the game. The fastest participant received 5 additional lotteries. Participants with the performance in top 5% received 4 additional lotteries. Participants with the performance between the top 5% and top 10% received 3 additional lotteries. Performance between the top 10% and top 15% was rewarded with 2 additional lotteries and finally, performance between top 15 and top 20% was rewarded with 1 additional lottery. Participant number 50 was selected from the lottery and paid out €30.

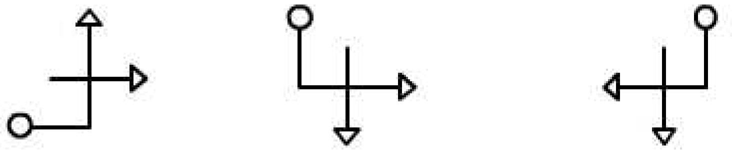
The main purpose of the new approach is to reduce the effect of other factors. Understanding how subjects estimate their own performance is key to eliminating potential biases. Larrick, Burson and Soll (2007) reported that when a person estimates his own performance he will do so by assessing the "memory of the recent performance, views of the self in that domain and general feelings about the self" (p. 79). By removing the "memory of the recent performance" I might be able to eliminate biases associated with memory such as availability heuristic, representativeness and misanthropy effect. I believe it is safe to assume that the majority of subjects are unfamiliar with cognitive games and thus have no memories and experiences that they can draw from.

To mitigate the illusion of knowledge, the game uses cognitive exercises as opposed to general knowledge questions. Illusion of knowledge is when confidence increases at a faster pace than the accuracy. Remember that illusion of knowledge was not present in the Line length test (Stankov and Crawford 1997) because there was no information about accuracy that the subjects could use. I argue that the same could be said about this cognitive game. To solve the game subject do not need to be knowledgeable in any particular domain. It is more a matter of intelligence and it is quite unlikely for subjects to be knowledge or experienced in playing cognitive games.

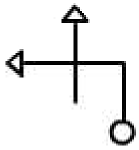
To control for possible idiosyncratic definition effects and ambiguity, the definition of a good performance in the task is well defined and unambiguous. A good final time indicates good performance. Even though the definition of a good time could be a bit vague and open to different interpretations, I believe it is safe to assume that subjects will at least have a general understanding of what is considered a good and what a bad time and will thus be able to correctly assess their performance. If overconfidence is observed in a subject who took 15 minutes to complete the exercise, it is most likely not due to his own definition of a good time. Taking 15 minutes to complete the exercise is undoubtedly a bad time.

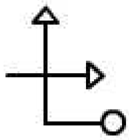
Here is an example of the cognitive exercise that was used in the experiment. Try to solve it and time yourself doing it. See how fast you can solve it. The correct answer of the exercise is revealed in the appendix B.

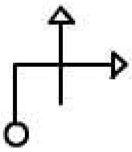
Example 1:

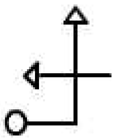


Which picture can be used to continue the series?


A.


B.


C.


D.

3.2 Part 2 - Measuring Overconfidence

After the game, subjects are randomly assigned into two groups – control group and the treatment group. Randomizations is achieved with Qualtrics randomization feature. Overconfidence was measured by using overestimation, overplacement and overprecision as overconfidence measures.

Overestimation

Time overestimation is the first overconfidence measure. Overestimation refers to rating your performance better than it actually was. In both treatments, subjects were asked to assess their performance time. Subjects' perceived time was determined by having participants choose their performance time from the choice list consisting of different times. The choice list presented to subjects was not incentivized and the same for both groups. Time overestimation was calculated by looking at the difference between perceived time and actual time. Perceived time that is shorter than the actual time indicates overestimation because the shorter the time, the better the final score. For instance, an overestimation of 100 indicates that a participant believed his time was 100 seconds faster than his actual time.

Overplacement

Subjects were asked to rank their performance relative to others using different methods in two treatments. Overplacement measure, constructed from introspective perceived rank in the control group, was compared to that elicited with incentivized choices in the treatment group to see whether using real incentives had an impact on overplacement.

In the control group subjects selected their perceived rank from the multiple choice list consisting of different rank options. This is the choice list that was presented to subjects in the control group:

- ❍ my performance was in the bottom 10%
- ❍ my performance was better than at least 10% of other participants
- ❍ my performance was better than at least 20% of other participants
- ❍ my performance was better than at least 30% of other participants
- ❍ my performance was better than at least 40% of other participants
- ❍ my performance was better than at least 50% of other participants
- ❍ my performance was better than at least 60% of other participants
- ❍ my performance was better than at least 70% of other participants
- ❍ my performance was better than at least 80% of other participants
- ❍ my performance was better than at least 90% of other participants

To understand how this questionnaire was utilized, imagine a participant choosing option 3, being better than 20% of other participants. This selection suggests that this participant believed he was better than at least 20% of other participants but not more than 30% because he did not choose option 4. The middle point of the two options was used as the perceived rank, in this case 25%. Perceived rank is then compared to the actual rank to construct the overplacement measure.

In the treatment group, the method used to elicit subjective beliefs with incentivized choices was the *exchangeability method* (Baillon 2008). This method consists of multiple questions. In each question, subjects face precisely two lotteries where they have to choose between different rank options. All questions are incentivized. Subjects were told that if they were selected to receive real payment, one of these questions would be randomly selected. They would be paid €30 if their actual performance fell in the interval specified by their chosen option. No actual money was paid out because the performance of the selected participant did not fall in the specified interval.

The following table shows the set of questions that was presented to subjects in the treatment group.

Table 1: exchangeability method: set of questions presented to subjects in the treatment group

	A	B
Question 1:	❍ win €30 if my performance was better than at least 10% of other participants.	❍ win €30 if my performance was better than at most 10% of other participants
Question 2:	❍ win €30 if my performance was better than at least 20% of other participants	❍ win €30 if my performance was better than at most 20% of other participants
Question 3:	❍ win €30 if my performance was better than at least 30% of other participants	❍ win €30 if my performance was better than at most 30% of other participants
Question 4:	❍ win €30 if my performance was better than at least 40% of other participants	❍ win €30 if my performance was better than at most 40% of other participants
Question 5:	❍ win €30 if my performance was better than at least 50% of other participants	❍ win €30 if my performance was better than at most 50% of other participants
Question 6:	❍ win €30 if my performance was better than at least 60% of other participants	❍ win €30 if my performance was better than at most 60% of other participants
Question 7:	❍ win €30 if my performance was better than at least 70% of other participants	❍ win €30 if my performance was better than at most 70% of other participants
Question 8:	❍ win €30 if my performance was better than at least 80% of other participants	❍ win €30 if my performance was better than at most 80% of other participants
Question 9:	❍ win €30 if my performance was better than at least 90% of other participants	❍ win €30 if my performance was better than at most 90% of other participants

In the first question, subjects had to choose between option A: receiving €30 if my performance can beat at least 10% of other participants (performance being better anywhere from 10%-100% of all participants) or option B: receiving €30 if my performance could beat at most 10% of other participants (performance being better anywhere from 0%-10% of all participants). The subject chooses the option that he finds more likely to win depending on his judgement of his performance relative to others. Subject had to make the same choice for all deciles. At some point, subjects switched from preferring option A to preferring option B. Subjects should switch only once. When they switch, they essentially put a cap on their perceived performance. For instance, imagine a subject selecting option B in the second choice question, selecting the option that his performance could beat at most 20% of other participants. It follows that he could only select option B in the third question, as option A in question 3 directly contradicts his answer from question 2. The middle point of where the subjects switched is taken as the point of indifference. For instance, imagine that a subject chooses option A in question 5: receiving €30 if my performance can beat at least 50% and then option B in question 6: receiving €30 if my performance can beat at most 60%. Based on his choices, we know that this subject believes his performance was better than 50% of other participants but not better than 60% of other participants. The middle point, 55% is used as the point of indifference and as that subject's perceived rank. It means that this subject believes his performance was better than 55% of other participants. Because this is his point of indifference he must believe that there is a 50% chance that his performance was actually better than 55% of other participants and a 50% chance that his performance was actually worse than 55% of other participants. Such an inference about the subject's belief relies on the assumption of additivity, which means that people's subjective probabilities of two complementary events must sum up to 100%. In the limitations section I will discuss whether this assumption is always satisfied but for now, let's imagine that additivity assumption holds.

The second overconfidence measure, *Performance overplacement*, was constructed from the perceived rank. Remember in the control group, perceived rank was elicited with a multiple choice list whereas in the treatment group, the exchangeability method was used. Overplacement was calculated for each participant by subtracting their actual percentage of participants beaten from their perceived percentage of participants beaten. A positive number indicates overplacement because it means that a participant believed his performance relative to others was better than it actually was.

In addition to giving subjects a financial incentive to state their beliefs, the main advantage of the exchangeability method is that it avoids ambiguity and risk biases, commonly associated with standard lottery methods. People might prefer to bet on a less risky prospect or a less ambiguous lottery. Ellsberg (1961) documented that people prefer known probabilities to unknown ones. This is a common problem with lottery methods where the point of indifference is obtained by directly comparing choices from two different sources. This is not the case with the exchangeability method where subjects compare choices from only one source (Baillon 2008). All rank choices in the exchangeability method are equally risky and ambiguous, which ensures risk and ambiguity preferences do not influence the results. Additionally, innovative approach of the exchangeability method ensures subjects answer the questions by telling the truth, because this is the best strategy to win the money. For those reasons, I considered the exchangeability method as the most appropriate method of measuring subjects' beliefs.

Overprecision

In the treatment group, a 75% confidence intervals of subjects' beliefs about ranking was elicited using two additional sets of questions, also using the exchangeability method. After answering the first set of questions, subjects then had to choose between new options to determine the new points of indifference. Choices from the first set determine which second and third set is later presented to the subjects. A point of indifference obtained from the first set is used as the upper boundary in the second set and as the lower boundary in the third set. An example of the second and third set and how they are connected to the first set is presented in Appendix C. Remember that the point of indifference in the first set splits the state-space into two equally likely events, because subjective probability of each of the two complementary events is 50%. Using the same logic, the points of indifference from the second and third sets now split the state-space into 4 equally likely events. Each of the 4 events has a 25% chance of occurring. Once those two new points of indifference are found, subjects' 75% confidence intervals can be formulated. Imagine we obtained a particular subject's second and third point of indifference of 52% and 62% respectively. It follows that this subject believes there is a 25% chance of his performance being worse than 52% of other participants and a 25% chance of his performance being better than 62% of other participants. This assumption allows us to construct that subject's 75% confidence intervals.

Performance overprecision is the third overconfidence measure. This measure was calculated by looking at subjects' 75% confidence intervals that were elicited with the exchangeability method in the treatment group. A score of 0 was assigned to subjects whose actual performance did not fall between their stated intervals and a score of 1 was assigned to subjects whose actual performance did fall between their stated intervals. If subjects were perfectly calibrated we would expect the average score to be 0.75. A lower score can be interpreted as subject being miscalibrated.

3.3 Hypotheses

First prediction concerns the effect of monetary incentives on elicited overplacement. The advantage of monetary incentives is that it gives subjects a financial incentive to state their true beliefs. However, if subjects are genuinely overconfident, monetary incentives should not influence their confidence levels because they truly believe that they are better than average. On the other hand, if subjects only appear overconfident because of social norms or because they self enhance and are thus not genuinely overconfident, then if they prefer money over self enhancing, incentivized choices might produce a difference in overplacement between the control and the treatment group. Thus, if subjects are genuinely overconfident then:

1. *There will be no difference in overplacement between the control group and the treatment group*.

Second prediction concerns the correlation between different overconfidence manifestations. As discussed earlier in the paper, a psychological bias related to genuine overconfidence cannot choose where to appear and it will influence self-perception across different domains. If people have a tendency to inflate their self-perception, then it does not make sense that a person would inflate their self-perception when evaluating his own performance, but would go in the complete opposite direction when comparing that performance to the others. Therefore, genuinely overconfident subjects will be overconfident regardless of which manifestation of overconfidence we are talking about. The second hypothesis therefore states:

2. *Subjects will be overconfident in all three measures of overconfidence: overestimation, overplacement and overprecision.*

Continuing with this line of thought, it follows that manifestations of overconfidence should also be related. When a person overestimates his performance, we would expect that person to also overplace it by an equally large margin. The third prediction states that:

3. *Different measures of overconfidence are correlated.*

3.4 Limitations

Methods used in this thesis to measure overconfidence have several limitations. The most obvious one is that subjects did not play the game in a controlled environment. The main purpose of the game was to limit the effects of other biases, but playing the game in a non-controlled environment introduces a host of other factors that could affect the results. A much better approach would be to have the subjects play the game in a lab or in some form of a controlled environment. Nevertheless, playing the game in such a setting does have some advantages. Since the game is anonymous and subjects play the game at home, experimenter demand effect should not be present. Furthermore, seeing other subjects in a lab might influence the results and how subjects rank themselves relative to others.

I argued that the exchangeability method is a better alternative to the lottery method because it avoids ambiguity and risk biases. However, the method is not without flaws. Baillon (2008) pointed out that the "method is based on the assumption that beliefs are additive" (p. 77). This means that when the event is split into two subevent, the probabilities of each subevent must sum up to one. Baillon (2008) showed that additivity can in some cases be violated. Cerroni et al. (2012) questioned the incentive compatibility of the method because of the use of chained questioned. This is a type of the elicitation technique where the answers to the first set of questions determine which set of questions is presented to the subjects next. In their study, they showed that the advantages of using monetary incentives are substantially diminished when using chained questions as the elicitation method. Furthermore, if subjects are aware of the chaining they can choose to answer questions strategically, in which case the elicited subjective probabilities are invalid. Incentive compatibility was also mentioned by Baillon (2008). He dealt with the problem by making subjects unaware of the chaining and the relationship between different questions. This principal was also applied in the design of this thesis's experiment. Subjects played the game and answered questions on their computer. The program was designed so that subjects could not see which questions and how many are coming next, so chaining was hidden in the design. However, since this was not a controlled environment, subject could have obtained information about chaining from other participants in which case their answers would not have reflected their true beliefs. Furthermore, Cerroni et al. (2012) also criticized the method by saying that the experimenter has to ask subject pointless questions that they have already ruled out in the previous sets of questions. Particularly this happens in the elicitation of the second and third point of indifference.

Further limitations of the methodology are related to the interpretation of overconfidence measures. Time overestimation was measured by asking participants about their perception of their final time. The interpretation of this measure is limited at best. Subjects could have checked the time on their computer or they could have timed it themselves.

Overprecision measure was recorded only with subjects in the treatment group. As a result, we cannot know if subjects in the control group are overconfident using this measure and what effects monetary incentives have on overprecision.

Different methods were used to obtain different overconfidence measures. For overestimation, subjects in both groups chose their time from the non-incentivized choice list. Performance overplacement in the control group was measured by having participants select their rank from the choice list and overplacement in the treatment group was measures with the exchangeability method. Overprecision was measured with the exchangeability method as well, but this measure of overconfidence was not recorded for the control group. Comparing overconfidence elicited with different techniques can be somewhat limited.

4. Results

4.1 Descriptive Statistics

I was able to collect 74 responses. Out of the 74, I only used 63 surveys. 11 surveys in total were eliminated. 8 surveys were eliminated because those participants did not finish the survey and the results were incomplete. 3 survey were eliminated because their results were inconclusive. 3 subjects, who chose their rank from the binary options of exchangeability method, switched from option A to option B more than once and I could thus not construct an overconfidence measure for those participants.

Three measures of overconfidence were used: performance overplacement, time overestimation, and performance overprecision. This section focuses mostly on performance overplacement and time overestimation. Overprecision is addressed at the end of this section.

Subjects in the experiment believed they were on average better than 61% of other participants. The median was 65%. Participants on average overplaced their performance by 11 percentage points. They also overestimated their total time. The average difference between perceived time and actual time was 230 seconds. Descriptive statistics are shown in table 2 and 3. One-sample

Wilcoxon test was performed to determine whether time overestimation and performance overplacement were significantly larger than 0, which would indicate overconfidence. Results were significant at a 0.05 level for performance overplacement and at a 0.01 level for time overestimation. This shows that participants were overconfident according to the overestimation and overplacement measure. Of course a time overestimation of say 10 seconds could hardly be regarded as overconfidence. Therefore, this measure was tested against 60 seconds to allow for a margin of error. The results still showed overconfidence at a 0.01 significance level. Even when increasing the margin to 120 seconds, overconfidence was still significant at the 0.01 level. Complete test statistics is presented in Appendix A, table 24.

Additional measure introduced in table 3 is *performance evaluation error* which is a variation of the overplacement measure. The overplacement measure used in this thesis, does not show by how much participants misplaced their performance. This is because overplacement is to a certain degree offset by underplacement. The absolute value of all performance overplacement data was taken and analysed in order to determine the error. This was done so that overplacement is not offset by underplacement. The average error of 26 percentage points and the median of 22 percentage points indicate that participants were not successful in assessing their performance relative to others. They overplaced themselves more than underplaced, but the error in both ways was quite large.

Table 2: descriptive statistics for all participants

All participants	Total time in s.	Perceived time in s.	% of participants beaten	Perceived % of participants beaten
average	716	486	50	61
median	657	450	50	65

Table 3: descriptive statistics – overconfidence measures

	Mean	Median	Minimum	Maximum	Std. Deviation
time overestimation in seconds	230	223	-130	761	204,643
performance overplacement in percentage points	11,03	7,54	-69,12	88,65	32,12
Performance evaluation error in percentage points	25,84	21,98	,24	88,65	21,83

There is considerable noise surrounding overconfidence measures. High standard deviations for all measures, average performance evaluation error of 26 percentage points as well as a substantial difference between a minimum and maximum, point to the likelihood of noisy data.

Figure 1 shows performance overplacement by ranks numbered from 1 to 63. Rank 1 is the top performing participant and rank 63 is the worst performing participant. Results show that poor performing participants were more overconfident than strong performing participants. In fact, most participants in the very top were underconfident. 8 of the top 10 performing participants displayed underconfidence, and 2 were able to correctly assess their performance. However, this does not come as a surprise, as the better the performance relative to others, the smaller the margin to be overconfident. The best performing participants had very little room to overplace their performance, as their performance was already at the very top. The exact opposite is true for the worst performing participants.

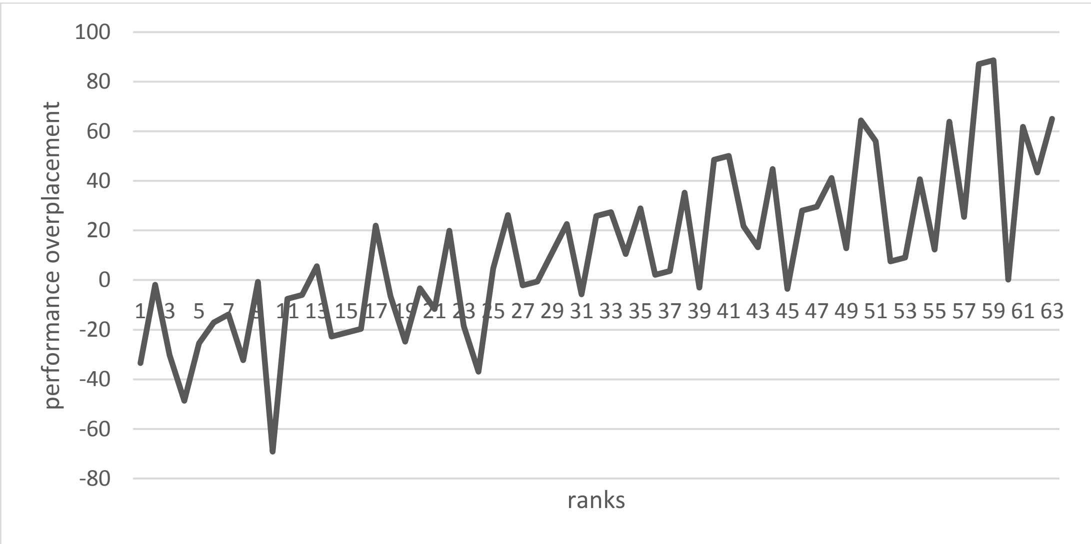


Figure 1: overplacement by ranks

4.2 Treatment effect

Descriptive statistics for the control group and the treatment group are shown in table 4 and 5. The median time was 667 seconds for control group and 657 seconds for the treatment group. Median time overestimation was 248 seconds and 196 seconds, respectively. Looking at the descriptive statistics, we can see that the performance in both groups was on average the same. The average participant in the control group was better than about 50% of other participants and so was the average participant in the treatment group. This shows that randomization was successful. The average perceived percentage of participants beaten was 64% in control group and 56% in the treatment group and the medians were 65% and 55%, respectively. The performance evaluation error was large in both groups, participants in control group on average misplaced their performance by 26 percentage points whereas participants in the treatment group did so by an average of 25 percentage points.

To test if there is a difference between performance overplacement in the control group and performance overplacement in the treatment group, I performed a Mann-Whitney U test. Results showed no significant difference between the control and the treatment group. The 2-tailed p value of a Mann-Whitney U test was 0.5. Full test statistics (table 10) is available in the Appendix A. There is also no significant difference in performance evaluation error between the groups. The p value of a Mann-Whitney U test is 0.4 (table 11 in the Appendix A). We cannot reject the null hypothesis that performance overplacement is the same in both samples. When given monetary incentives participants were not more successful in placing their performance relative to others. It follows that monetary incentives did not have a large enough effect to significantly influence participants' assessment of their own performance. Further interpretation of this result can be read in the discussion part of the thesis.

Table 4: descriptive statistics - control group

Treatment 0	Total time	Perceived time	Time overestimation	% of participants beaten	Perceived % of participants beaten	Performance overplacement	Performance evaluation error	N
average	737	450	287	49	64	15	26	29
median	667	450	248	49	65	4	20	

Table 5: Descriptive statistics - treatment group

Treatment 1	Total time	Perceived time	Time overestimation	% of participants beaten	Perceived % of participants beaten	Performance overplacement	Performance evaluation error	N
average	691	528	163	50	56	6	25	31
median	657	450	196	51	55	12	25	

4.3 Gender effect

This thesis also studied the effect of gender on overconfidence. Descriptive statistics for men are displayed in table 6 and 7 and for women in table 8 and 9. Significant difference at the 0.01 level in performance overplacement between men and women was found using the Mann-Whitney U test. Statistics table is presented in the Appendix A, table 12. Results show that men were considerably more overconfident than women. In fact, women were not overconfident at all, with the average performance overplacement being less than 1 percentage point. Conversely, the average performance overplacement in the men sample was 22 percentage points. On a population level, women were able to almost perfectly assess their performance relative to others. They believed their performance was on average better than 53% of other participants, which is almost exactly on par with the actual percentage. Men on the other hand, were arguably quite misguided and noticeably overconfident as they believed their performance was on average better than 68% of other participants, with the actual number being only 46% of other participants. Moreover, women were not only less overconfident than men, they were also more successful than men in evaluating their performance relative to others. The average error made in evaluating performance relative to others was 34 percentage points for men and only 17 percentage points for women. The difference is significant at the 0.01 level, tested with the Mann-Whitney U test (table 13, Appendix A). Furthermore, standard deviation for overconfidence measures was smaller for women than for men, which points to women being more consistent than men in evaluating their performance.

There was no significant difference in time overestimation between men and women (table 14, appendix A). This is intriguing because it means that men were overconfident in both measures (overestimation and overplacement), whereas women were overconfident only in one measure (overestimation).

Table 6: Descriptive Statistics - Men

MEN	Total time	Perceived time	% of participants beaten	Perceived % of participants beaten	N
average	750	500	46	68	31
median	706	510	41	75	

Table 7: descriptive statistics men – overconfidence measures

MEN	N	Mean	Median	Minimum	Maximum	Std. Deviation
time overestimation in seconds	31	250	223	-130	761	218
performance overplacement in percentage points	31	22,11	26,27	-69,12	88,65	36,85
Performance evaluation error in percentage points	31	34,46	28,97	,71	88,65	25,23

Table 8: Descriptive statistics - Women

WOMEN	Total time	Perceived time	% of participants beaten	Perceived % of participant beaten	N
average	682	472	53	53	32
median	652	450	55	55	

Table 9: descriptive statistics women – overconfidence measures

WOMEN	N	Mean	Median	Minimum	Maximum	Std. Deviation
time overestimation in seconds	32	210	216	-121	637	191,54
performance overplacement in percentage points	32	,29	-,15	-48,65	50,08	22,51
Performance evaluation error in percentage points	32	17,49	14,92	,24	50,08	13,82

A two by two between subjects Anova was performed to test for interaction effects. The results have already shown us that there is no significant difference in overplacement between control and treatment group and a significant difference at the 0.01 level in overplacement between men and women. Anova was performed to test for treatment effect within the gender group. Not surprisingly, there was a significant effect of gender on overplacement at the 0.01 level. This was already found using the Man-Whitney U test. However, Anova test revealed there was no interaction effect. Treatment*gender was not significant with a p value of 0.11. Results suggest that incentivizing men and women did not significantly affect their overconfidence levels. The same test was conducted for the performance evaluation error. Results were the same. A significant effect of gender but no significant effect of treatment and treatment*gender.

Results of the Anova tests as well as descriptive statistics for within gender and treatment can be found in tables 15 to 18 in the Appendix A.

4.4 Correlation between time overestimation and performance overplacement

Further analysis of the data reveals a significant correlation at the 0.01 level between time overestimation and performance overplacement. Bivariate Pearson correlation analysis was conducted to determine the correlation. A positive correlation of 0.509 indicates that the more the participants overestimated their time, the more they also overplaced their performance. The correlation was more pronounced in the control group. The correlation of 0.58 is significant at the 0.01 level. The correlation of 0.36 in the treatment group is only significant at the 0.05 level. Moreover, the correlation is also more pronounced with men. The correlation of 0.64 in the men sample is significant at the 0.01 level. Because women were not overconfident in overplacement measure but were overconfident in overestimation measure, the correlation is not significant in the women sample. Complete statistics tables are available in the Appendix A, tables 19 to 23.

4.5 Overprecision

Overprecision was tested using the 75% confidence intervals elicited with the exchangeability method. Confidence intervals were obtained only from the subjects in the treatment group, therefore only 29 responses were used to test for overprecision. A score of 0 was assigned to subjects whose actual performance did not fall between their stated intervals. Subjects whose actual performance did fall between their stated intervals were assigned a score of 1. We would expect an average score of 0.75 if subjects were perfectly calibrated. A lower score would indicate overprecision. The average score was 0.24. The binominal test confirmed that this score is indeed statistically significantly different than 0.75 at the 0.01 level. This means that subjects were miscalibrated. Test statistics table is presented in the Appendix A, table 25. The average range, calculated by deducting the lower from the upper boundary, was 11.2 percentage points and the median range was only 7 percentage points. The median and the average performance evaluation error in the treatment group was 25 percentage points which shows that subjects' stated confidence intervals were too narrow.

5. Discussion

The first hypothesis predicted no difference in overplacement between the control group and the treatment group, if subjects were genuinely overconfident. Statistical tests revealed no significant difference between the control and the treatment group, therefore we cannot reject the null hypothesis. When given monetary incentives, participants were not more successful in placing their performance relative to others. It follows that monetary incentives did not have a large enough effect to significantly influence participants' assessment of their own performance. One possible interpretation is that participants were in fact genuinely overconfident. However, we must be careful not to jump to conclusions too quickly. The critic of Cerroni et al. (2012) about the use of chained questions might be a possible explanation of why there was no statistical difference. Furthermore, I hypothesized that if subjects were not genuinely overconfident, monetary incentives should produce a difference between the control and treatment group because subjects prefer money to self-enhancing and social norms. It is possible that this is not true. It might be that subjects in fact prefer self-enhancing, and this is why there was no statistical difference. I did not measure the utility of self-enhancement and it is possible that monetary incentives used in this experiment were too small.

Results also support our second hypothesis that stated subjects would be overconfident in all three measures of overconfidence. Subjects overestimated their time, overplaced their performance relative to others and provided too narrow confidence intervals which resulted in miscalibration. However, time overestimation measure is perhaps not very informative. Subjects could have checked the time on their computer or they could have timed it themselves, so this measure is hard to interpret. However, remember that the final time consists of recorded time plus additional two minutes for every incorrect answer. This suggests that participants did in some way overestimate their performance. This could be because they overestimated the number of questions that they answered correctly, or because they overestimated the time they needed to finish the game. Nevertheless, it is unclear how that overestimation came about. The outcome that participants were overconfident in all three measures can only be said about the treatment group, as overprecision was not tested with the control group. Additional limitation is that different methods were used to elicit overconfidence, so even though we cannot reject the null hypothesis, we also cannot claim that this result confirms genuine overconfidence.

The third hypothesis predicted a correlation between different overconfidence manifestations because they are all a product of the same bias. A higher overestimation should result in higher overprecision and vice-versa. While this was found in the whole sample, in both the control and the treatment group, and also in the men sample, the correlation was not significant in the women sample. Women did overestimate their time but did not overplace their performance. Does this mean that we can reject our third hypothesis? Because the correlation was pronounced in all but the women sample, this question is hard to answer. It is possible that different overconfidence manifestations do not materialize in the same fashion but are still a product of the same bias. On the other hand, it is also possible that different manifestations of overconfidence have a different underlying cause, as suggested by Moore et al. (2008) and Klayman (1999). Further research is required to determine the connection between different overconfidence manifestations.

Results also show that men were more overconfident than women. This is in line with the literature. Many studies have reported men being more overconfident, more prone to risks and more prone to the self-attribution bias than women (Barber and Odean 2001).

Finally, I would like to conclude by giving my opinion on the subject of overconfidence. Is overconfidence bad? Many researchers say yes. For instance, Larrick et al. (2007) say that overconfidence can lead to unwise choices. In financial markets, Barber and Odean (2000) documented that overconfident investors hold under-diversified portfolios and engage in excessive trading. Malmendier and Tate (2005) reported that overconfident CEOs are more sensitive to cash flows and that they will underinvest with insufficient internal funds and overinvest with sufficient internal funds. Furthermore, they suggested that overconfident CEO will engage in more value destroying mergers. Studying the causes for overconfidence is therefore extremely important. If we want to mitigate overconfidence and help people make better choices, we need to know how overconfidence effect is created. Different biases warrant different interventions.

But there is also the other side. Hirshleifer, Low, & Teoh (2012) documented that overconfident CEO are better innovators and can better translate growth opportunities into firm value. My belief is that a healthy dose of overconfidence is a good thing, particularly when it comes to innovation and competition. Imagine a game of poker. If all players were able to correctly assess their skills, there would be no one left to play the game. The bottom 50% of players would immediately drop out because they would conclude that they have no chance of winning the money. After they drop out, players would have to reassess their relative position. Again, the bottom 50% would drop out. This would continue until only one player was left. And then even he would stop playing because there would be no one to play him and the game would eventually die out. Sports, entrepreneurship and different types of competition would most likely suffer the same. While too much overconfidence can definitely lead to unwise decisions, moderate overconfidence might be one of the crucial parts of the competition engine.

6. Conclusion

In conclusion, can we say that we found genuine overconfidence in this experiment? Results from the experiment support several of the genuine overconfidence hypotheses. However, the limitations of this study prevent us from comfortably making the conclusion that overconfidence is indeed a psychological bias. Further research is required to verify the existence of genuine overconfidence. I would recommend future researchers to use the new approach presented in this thesis and improve it by studying its limitations. The purpose of this paper however was not only to measure genuine overconfidence but also to question the general perception of what overconfidence is. The purpose was to research other possible sources that could cause subjects to appear overconfident, to contribute to our understanding of overconfidence and to provide a different, perhaps better approach of measuring it. Even though genuine overconfidence per-se could not be confirmed, I believe this thesis made its contribution to the cause of better understanding overconfidence.

Last but not least, I would like to conclude with one final thought. Does overconfidence always sink the ship as Oscar Wilde put it? Many ships have been sunk in the deep and turbulent overconfidence waters but some have made it through. Without captains confident enough to engage in that endeavour, we would never have any ships sailing the sea.

"*The moment you doubt whether you can fly, you cease for ever to be able to do it*." J.M. Barrie: Peter Pan.

References

1. Baillon, A. (2008). Eliciting subjective probabilities through exchangeable events: An advantage and a limitation. *Decision Analysis*, *5*(2), 76-87.
2. Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance*, *55*(2), 773-806.
3. Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly journal of Economics*, 261-292.
4. Beasley, R. (2016). Dissonance and decision-making mistakes in the age of risk. *Journal of European Public Policy*, *23*(5), 771-787.
5. Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *The American economic review*, *96*(5), 1652-1678.
6. Ben-David, I., Graham, J. R., & Harvey, C. R. (2010). *Managerial miscalibration* (No. w16215). National Bureau of Economic Research.
7. Ben-David, I., Graham, J. R., & Harvey, C. R. (2010). *Managerial miscalibration* (No. w16215). National Bureau of Economic Research.
8. Benoît, J. P., Dubra, J., & Moore, D. A. (2015). Does the better‐than‐average effect show that people are overconfident?: two experiments. *Journal of the European Economic Association*, *13*(2), 293-329.
9. Bhandari, G., & Deaves, R. (2006). The demographics of overconfidence. *The Journal of Behavioral Finance*, *7*(1), 5-11.
10. Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination.*Organizational Behavior and Human Decision Processes*, *65*(3), 212-219.
11. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability.*Monthly weather review*, *78*(1), 1-3.
12. Carson, A. D. (1998). The relation of self-reported abilities to aptitude test scores: A replication and extension. *Journal of Vocational Behavior*, *53*(3), 353-371.
13. Cerroni, S., Notaro, S., & Shaw, W. D. (2012). Eliciting and estimating valid subjective probabilities: An experimental investigation of the exchangeability method. *Journal of Economic Behavior & Organization*, *84*(1), 201-215.
14. Cerroni, S., Notaro, S., & Shaw, W. D. (2012, June). The validity of risk estimates elicited via the Exchangeability Method: An experimental investigation of consumers' perceived health risks. In *Congress Papers* (No. 124100). Italian Association of Agricultural and Applied Economics (AIEAA).
15. Creswell, J. D., Lam, S., Stanton, A. L., Taylor, S. E., Bower, J. E., & Sherman, D. K. (2007). Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing?. *Personality and Social Psychology Bulletin*, *33*(2), 238-250.
16. Eisenbach, T. M., & Schmalz, M. C. (2015). Anxiety, overconfidence, and excessive risk taking. *FRB of New York Staff Report*, (711).
17. Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 643-669.
18. Emich, K. J. (2014). But consider the alternative: The influence of positive affect on overconfidence. *Cognition and Emotion*, *28*(8), 1382-1397.
19. Glaser, M., Langer, T., & Weber, M. (2013). True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making*, *26*(5), 405-417.
20. Grubb, M. D. (2015). Overconfident consumers in the marketplace. *The Journal of Economic Perspectives*, *29*(4), 9-35.
21. Hepper, E. G., Gramzow, R. H., & Sedikides, C. (2010). Individual differences in self‐enhancement and self‐protection strategies: An integrative analysis.*Journal of personality*, *78*(2), 781-814.
22. Hilton, D., Regner, I., Cabantous, L., Charalambides, L., & Vautier, S. (2011). Do positive illusions predict overconfidence in judgment? A test using interval production and probability evaluation measures of miscalibration. *Journal of Behavioral Decision Making*, *24*(2), 117-139.
23. Hirshleifer, D., Low, A., & Teoh, S. H. (2012). Are overconfident CEOs better innovators?. *The Journal of Finance*, *67*(4), 1457-1498.

24. Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, *10*(3), 269-278.
25. Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational behavior and human decision processes*, *79*(3), 216-247.
26. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.
27. Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76-94.
28. Latham, C. (June 22, 2007). Pattern Recognition Brain Teaser – The Empty Triangle. Retrieved May 25, 2016, from http://sharpbrains.com/blog/2007/06/22/pattern-recognition-brain-teaser-the-empty-triangle/
29. Lee, J. (2012). Trait desirability and cultural difference in the better-than-average effect. *Asian Journal of Social Psychology*, *15*(4), 261-272.
30. Ma, N. (2008). Overconfidence—its theory and implications in dispute resolution. *ADR Bulletin*, *10*(4), 4.
31. Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. *The journal of finance*, *60*(6), 2661-2700.
32. Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence.*Psychological review*, *115*(2), 502.
33. Murad, Z., Sefton, M., & Starmer, C. (2016). How do risk attitudes affect measured confidence?. *Journal of Risk and Uncertainty*, *52*(1), 21-46.
34. Pattern recognition. Intelligencetest.com. Retrieved May 25, 2016, from http://www.intelligencetest.com/index.htm
35. Pattern recognition. Iqtestexperts.com. Retrieved Mayy 25, 2016, from http://www.iqtestexperts.com/pattern-recognition-sample.php
36. Pfeffer, J., Cialdini, R. B., Hanna, B., & Knopoff, K. (1998). Faith in supervision and the self-enhancement bias: Two psychological reasons why managers don't empower workers. *Basic and Applied Social Psychology*, *20*(4), 313-321.
37. Recognizing Visual Patterns. *Brilliant.org*. Retrieved May 25, 2016, from https://brilliant.org/wiki/pattern-recognition-visual-easy-2/
38. Roy, M. M., & Liersch, M. J. (2013). I am a better driver than you think: examining self-enhancement for driving ability. *Journal of applied social psychology*, *43*(8), 1648-1659.
39. Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought.*Perspectives on Psychological Science*, *3*(2), 102-116.
40. Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*(2), 117-137.
41. Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, *25*(2), 93-109.
42. Steele, C. M., Spencer, S. J., & Lynch, M. (1993). Self-image resilience and dissonance: the role of affirmational resources. *Journal of personality and social psychology*, *64*(6), 885.
43. Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers?. *Acta psychologica*, *47*(2), 143-148.
44. Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology*, *69*(6), 1120.
45. Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.
46. Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.
47. Weinberg, B. A. (2009). A model of overconfidence. *Pacific Economic Review*,*14*(4), 502-515.
48. What Comes Next?. *Brilliant.org*. Retrieved May 25, 2016, from https://brilliant.org/wiki/pattern-recognition-what-comes-next/

49. Ybarra, O. (1999). Misanthropic person memory when the need to self-enhance is absent. *Personality and Social Psychology Bulletin*, *25*(2), 261-269.

Appendix

A-Tables

Table 10: Mann-Whitney U test - dependent variable: performance overplacement, grouping variable: treatment

Ranks

	Treatment	N	Mean Rank	Sum of Ranks
Performance overplacement	control	34	33,38	1135,00
	treatment	29	30,38	881,00
	Total	63		

Test Statistics

Mann-Whitney U	446,000
Z	-,648
Asymp. Sig. (2-tailed)	,517
Exact Sig. (2-tailed)	,524

Table 11: Mann-Whitney U test - dependent variable: performance evaluation error, grouping variable: treatment

Ranks

	treatment	N	Mean Rank	Sum of Ranks
Performance evaluation error	control	34	30,22	1027,50
	treatment	29	34,09	988,50
	Total	63		

Test Statistics

Mann-Whitney U	432,500
Z	-,834
Asymp. Sig. (2-tailed)	,404
Exact Sig. (2-tailed)	,409

Table 12: Mann-Whitney U test – dependent variable: performance overplacement; grouping variable: gender

Ranks

	gender	N	Mean Rank	Sum of Ranks
performance overplacement	women	32	25,88	828,00
	men	31	38,32	1188,00
	Total	63		

Test Statistics

Mann-Whitney U	300,000
Z	-2,695
Asymp. Sig. (2-tailed)	,007
Exact Sig. (2-tailed)	,007

Table 13: Mann-Whitney U test – dependent variable: performance evaluation error; grouping variable: gender

Ranks

	gender	N	Mean Rank	Sum of Ranks
Performance evaluation error	women	32	25,77	824,50
	men	31	38,44	1191,50
	Total	63		

Test Statistics

Mann-Whitney U	296,500
Z	-2,743
Asymp. Sig. (2-tailed)	,006
Exact Sig. (2-tailed)	,005

Table 14: Mann-Whitney U test – dependent variable: time overestimation; grouping variable: gender

Ranks

	gender	N	Mean Rank	Sum of Ranks
Time overestimation	women	32	30,91	989,00
	men	31	33,13	1027,00
	Total	63		

Test Statistics

Mann-Whitney U	461,00
Z	-,481
Asymp. Sig. (2-tailed)	,630
Exact Sig. (2-tailed)	,635

Table 15: descriptive statistics - dependent variable: performance overplacement; grouping variable: gender, treatment

Descriptive Statistics

Dependent Variable: performance overplacement

Treatment	Gender	Mean	Std. Deviation	N
Control	Women	-2,16	22,82	16
	Men	30,39	35,16	18
	Total	15,07	33,85	34
Treatment	Women	2,76	22,66	16
	Men	10,65	37,40	13
	Total	6,30	29,84	29
Total	Women	,30	22,51	32
	Men	22,11	36,86	31
	Total	11,03	32,12	63

Table 16: two way ANOVA (gender, treatment) for the depended variable: performance overplacement

Tests of Between-Subjects Effects

Dependent Variable: performance overplacement

Source	Sum of Squares	df	Mean Square	F	Sig.
treatment	852,337	1	852,337	,943	,335
gender	6353,456	1	6353,456	7,029	,010
treatment * gender	2360,930	1	2360,930	2,612	,111

Table 17: descriptive statistics - dependent variable performance evaluation error; grouping variable: gender, treatment

Descriptive Statistics

Dependent Variable: Performance evaluation error

treatment	gender	Mean	Std. Deviation	N
Control	Women	16,10	15,80	16
	Men	35,50	29,66	18
	Total	26,36	25,75	34
Treatment	Women	18,88	11,87	16
	Men	33,02	18,45	13
	Total	25,22	16,51	29
Total	Women	17,49	13,82	32
	Men	34,46	25,22	31
	Total	25,84	21,83	63

Table 18: two way ANOVA (gender, treatment) for the depended variable: performance evaluation error

Tests of Between-Subjects Effects

Dependent Variable: Performance evaluation error

Source	Sum of Squares	df	Mean Square	F	Sig.
treatment	,390	1	,390	,001	,976
gender	4370,192	1	4370,192	10,353	,002
treatment * gender	108,204	1	108,204	,256	,615

Table 19: Time overestimation - performance overplacement correlation

All participants		time overestimation	performance overplacement
time overestimation	Pearson Correlation	1	,509**
	Sig. (2-tailed)		,000
	N	63	63
performance overplacement	Pearson Correlation	,509**	1
	Sig. (2-tailed)	,000	
	N	63	64

***. Correlation is significant at the 0.01 level (2-tailed).*

Table 20: Time overestimation - performance overplacement correlation (men sample)

Men		time overestimation	performance overplacement
time overestimation	Pearson Correlation	1	,647**
	Sig. (2-tailed)		,000
	N	31	31
performance overplacement	Pearson Correlation	,647**	1
	Sig. (2-tailed)	,000	
	N	31	31

***. Correlation is significant at the 0.01 level (2-tailed).*

Table 21: Time overestimation - performance overplacement correlation (women sample)

Women		time overestimation	performance overplacement
time overestimation	Pearson Correlation	1	,278
	Sig. (2-tailed)		,123
	N	32	32
performance overplacement	Pearson Correlation	,278	1
	Sig. (2-tailed)	,123	
	N	32	32

Table 22: Time overestimation - performance overplacement correlation (treatment group)

Treatment group		time overestimation	performance overestimation
time overestimation	Pearson Correlation	1	,358*
	Sig. (2-tailed)		,048
	N	31	31
performance overplacement	Pearson Correlation	,358*	1
	Sig. (2-tailed)	,048	
	N	31	31

*. *Correlation is significant at the 0.05 level (2-tailed).*

Table 23: Time overestimation - performance overplacement correlation (control group)

Control group		time overestimation	performance overestimation
time overestimation	Pearson Correlation	1	,580**
	Sig. (2-tailed)		,000
	N	36	36
performance overplacement	Pearson Correlation	,580**	1
	Sig. (2-tailed)	,000	
	N	36	36

**. *Correlation is significant at the 0.01 level (2-tailed).*

Table 24: Hypotheses test summary

Null Hypothesis	Test	Sig.	Decision
The median of performance overplacement equals 0	One-Sample Wilcoxon Signed Rank Test	0.028	Reject the null hypothesis
The median of time overplacement equals 0	One-Sample Wilcoxon Signed Rank Test	0.000	Reject the null hypothesis
The median of time overplacement equals 60	One-Sample Wilcoxon Signed Rank Test	0.000	Reject the null hypothesis
The median of time overplacement equals 120	One-Sample Wilcoxon Signed Rank Test	0.002	Reject the null hypothesis

Table 25: Binominal test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)	Exact Sig. (1-tailed)
in the range	Yes	1	7	,24	,75	,000	,000
	No	0	22	,76			
	Total		29	1,00			

B –result of the cognitive exercise

The correct answer of the exercise is option A.

C – Example of a set of questions used to elicit second and third point of indifference

Note: the first elicited point of indifference was in this case 55%.

Instructions: Read the last set of questions carefully. It might seem like you have answered some of them before, but they are a little bit different and the answers are important for my thesis. One of the question could also be randomly selected for real payment so give it your best try. In each of the following questions you will need to choose between option A and option

B. You will be paid 30 euros if your actual performance falls in the interval specified by your chosen option.

Table 26: set of questions used to elicit the second point of indifference

Question 1:	A ❍ Win €30 if my performance was better than at least 5% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 5% of other participants
Question 2:	A ❍ Win €30 if My performance was better than at least 15% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 15% of other participants
Question 3:	A ❍ Win €30 if my performance was better than at least 30% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 30% of other participants
Question 4:	A ❍ Win €30 if my performance was better than at least 40% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 40% of other participants
Question 5:	A ❍ Win €30 if my performance was better than at least 45% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 45% of other participants
Question 6:	A ❍ Win €30 if my performance was better than at least 48% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 48% of other participants
Question 7:	A ❍ Win €30 if my performance was better than at least 50% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 50% of other participants
Question 8:	A ❍ Win €30 if my performance was better than at least 52% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 52% of other participants
Question 9:	A ❍ Win €30 if my performance was better than at least 54% and at most 55% of participants	B ❍ Win €30 if my performance was better than at most 54% of other participants

Table 27: set of questions used to elicit the third point of indifference

Question 1:	A ❍ Win €30 if my performance was better than at least 57% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 57% of participants
Question 2:	A ❍ Win €30 if my performance was better than at least 59% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 59% of participants
Question 3:	A ❍ Win €30 if my performance was better than at least 61% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 61% of participants
Question 4:	A ❍ Win €30 if my performance was better than at least 64% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 64% of participants
Question 5:	A ❍ Win €30 if my performance was better than at least 70% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 70% of participants
Question 6:	A ❍ Win €30 if my performance was better than at least 80% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 80% of participants
Question 7:	A ❍ Win €30 if my performance was better than at least 90% of other participants	B ❍ Win €30 if my performance was better than at least 55% and at most 90% of participants