UiT

NORGES ARKTISKE UNIVERSITET

Handelshøgskolen

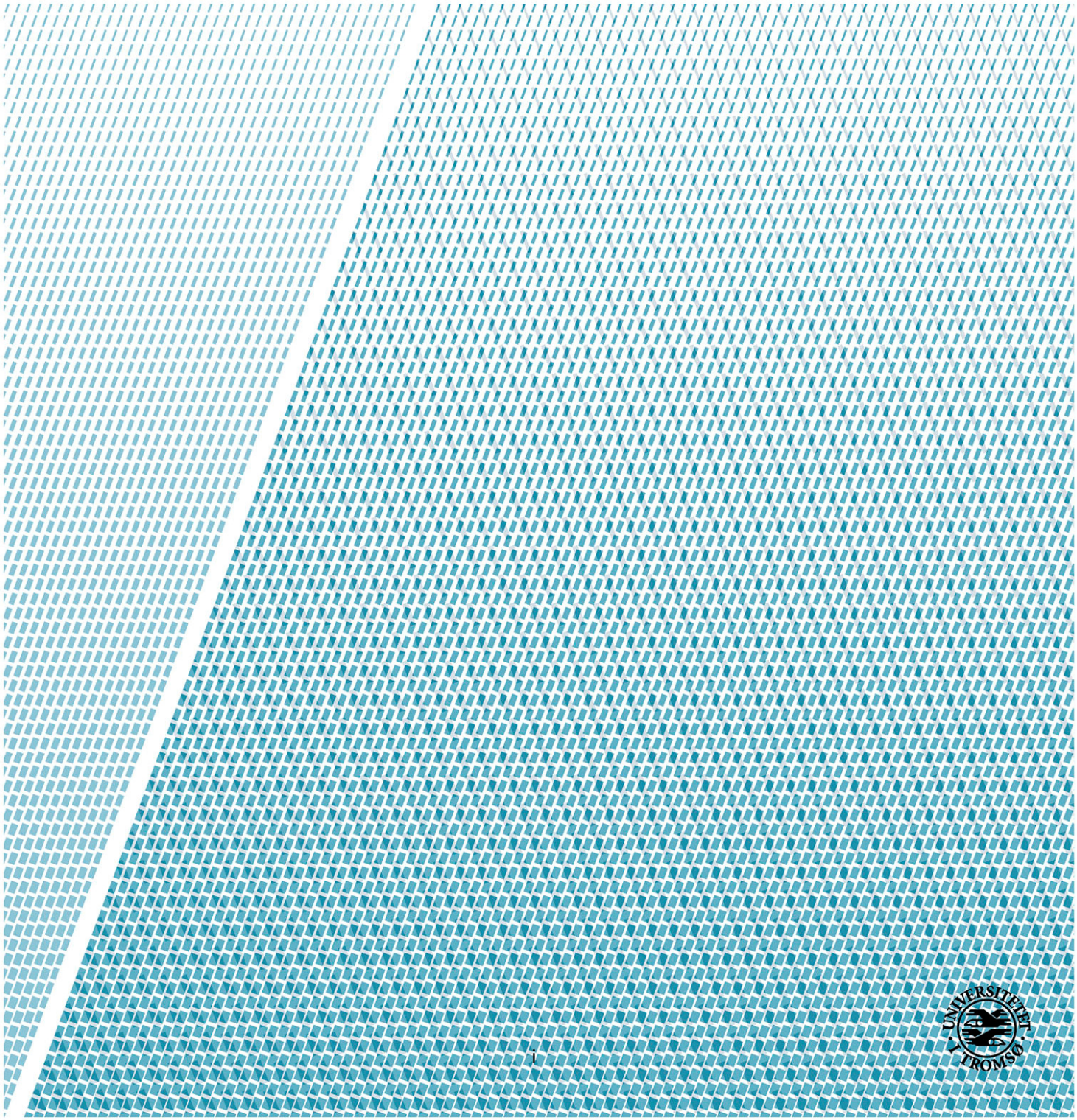
The Lunch effect

Can it result in biased grades at universities?

—

Johan Birkelund

Master thesis in economics, December 2014



i Preface

My time at the University of Tromsø has been interesting to say the least. I started studying economics by pure chance, as it was the first open study program I saw, one early august day in 2009. Although my study group with Truls, Rune and Kristian did not last throughout my time at the university, it made my days as a bachelor student truly memorable. Stian and Bjørn made the master years equally as good, with many late nights consisting of Cola Free, Kuhn-Tuckers, powerlifting videos and differential equations.

I am also glad to have reconnected with Eirik Heen, who I had not seen since high school several years prior to starting at the University. Lucky for me, it turned out we shared interest in several fields of research, which resulted in hours of interesting academic talks, and a memorable trip to NHH. On that note, I would also like to thank The Choice Lab at NHH, for accepting me as a visiting student for two Ph.D. courses. Of all topics in economics, behavioral is my very favorite, and I hope I can continue working in the field, after my master's degree is done.

I would also thank to my thesis advisor, Professor Stein Østbye, for giving good lectures, and convincing me to stay within economics.

Finally yet most importantly, I would like to give special thanks to Ida Harviken for letting me be a total geek! You are the best!

ii Index

iii Table index

iv Figure index

v Summary

Examining exam results from an oral exam at a Norwegian university, reveals that there may be supporting evidence for the existence of a lunch-effect. Results suggests that censors are making the “easy” choices right before lunch, as compared with right after. The results are both a support to Danziger et al. (2011a) and to the existing literature on biased grading in general.

Keywords: Ego-depletions, expert decision makers, fairness, grading

1. Introduction

During the last decades, a growing body of literature has emerged from behavioral economics. Economists have come to realize that the economic decisions of humans do not always follow the model of rational agents, also known as *homo economicus*. We are prone to several cognitive biases, and make our decisions not truly based on expected utility and revealed preferences. One of the fields within decision-making consists of papers on biased exam grading. In the case of grading, censors, which are considered experts on their field, are also prone to errors. It seems like students do not get grades based solely on their performance at their exam, but things like ethnicity, gender, halo and contrast effects play a part when the grade is determined (Fleming, 1999). These exogenous factors should have no effect on the grade, which is supposed to be based solely on the performance of the student. Even in the case of judicial decisions, there has been found effects from exogenous factors. Danziger et al. (2011a) tested the common caricature of "what the judge ate for breakfast", by examining data consisting of 1,112 judicial rulings, collected over a 10 month period in Israel. To their surprise, they found that the percentage of favorable rulings dropped gradually from about 65% to nearly zero within a decision session, and returned to 65% after a food break before reducing gradually again to about zero. This effect has been coined *"The lunch effect"*.

In Danziger et al.'s study, how long it had been since the judges had eaten, should have had no effect on the judicial decision. However, since meal breaks had a bearing on legal decisions, it's plausible to think that this could also be affecting decisions of "less importance" like the grading of student exams. If this is the case, it is reasonable to assume that this bias has serious implications in nearly all areas of decision-making.

Inspired by Danziger et al. (2011a)'s paper, which will be discussed thoroughly in part 3 of the paper, the current research seeks to examine the exam results from the fall semester 2012 at the University of Tromsø.

As for the literature on biased grading, as most of the papers, if not all, are based on high school grading. At least to my knowledge, no papers exist on biased exam grades at university level of education.

The way oral exams are conducted serves as a nice setup for a quasi-field experiment. The subjects of this experiment is the censors, who evaluates the student and give them their grade shortly after the students have given their examination. The treatments in this experiment is simply if the grade is set before or after lunch. The focus of the current research is whether the lunch-break affects the grade or not. A complete causal effect may be difficult to state given the information available, but it should be sufficient to detect a tendency in the grades given.

If I find any evidence of this effect by looking at exam results, it would in addition to providing support for Danziger et al. (2011a) and Linder et al. (2014)'s results, also be a contribution to the existing literature on biased grading, which to my knowledge, up until now do not have any papers from university level grading.

The thesis is composed in the following way. Part 1 consists of this introduction, and some implications of biased grades. Part 2 discusses the evolution of economics from traditional economics with strict assumptions about human behavior, via behavioral economics, and ending in experimental economics. Part 3 is a literature review of biased grading, ego-depletion and a more thorough look at Danziger et al. and Linder et al.'s papers on the lunch effect. Part 4 explains my data, method of analysis, hypothesis and results. Part 5 provides a discussion of the results presented, and part 6 is my concluding remarks.

1.1 Implications of biased grades

Empirical data show that more females than males enroll in university studies, where they are predominant in arts and humanities. Males, however is predominant in natural science fields. There is also common knowledge about a significant wage gap between men and women (Weichselbaumer and Winter-Ebmer, 2005). While there are several possible explanations for the differences in wages, one of the explanations may be the different educational choices made by men and women.

Based on the principal/agent cheap talk models by (Bénabou and Tirole (2000), Benabou and Tirole, 2003) where the principal is more informed about the agents abilities than the agent, and uses incentives to invoke over or underconfidence on the agent, Mechtenberg (2009) forms a unified model to provide an explanation to the gender differences in grading, wages

and enrollment at universities. Her explanation suggests that there is a bias in school grading, which results in these differences. The model she presents utilizes principals (teachers) and agents (students), where teachers can send good or bad signals (grades) to the students. She claims that teachers' grades are biased, because teachers use grades to signal if they like or dislike students' attitude. Mechtenberg also argues that men and women seem to put different meanings in these signals, leading to asymmetric information in the meaning of the grades between the teacher and the student. Women tend to interpret bad grades in humanities as bad messages of their attitude; rather than interpret it as a lack in abilities, but when receiving good grades, they interpret them as being skilled. When it comes to mathematics, women seem to interpret a bad grade as lack of skill and a good grade as a sign of good attitude. This in turn results in girls being overconfident in humanities, and underconfident in mathematics, leading them to apply for humanities studies rather than mathematics and science studies at universities. While this is the result for women, men have the reverse belief, leaving them to be overconfident in mathematics and underconfident in humanities.

If grades are biased, the result may be that individuals choose education, which is not suitable to their talents. This in turn can lead to larger wage inequality. While Mechtenber addresses biased grading at high school level, at universities, biased grading may have similar sorting implications causing inefficient selection for future studies and career tracks. In addition, biased grading could have direct monetary implications for the students. Most of the students at the University of Tromsø (and Norway in general) receive financial support for conducting their studies by Statens lånekasse for utdanning. Students receive a loan, and if they get a passing grade on their final exam, a part of the loan will be transformed to a scholarship of roughly $1000 per course.

2. Methodology

The first part of this chapter will briefly address traditional economic theory of decision-making and how traditional economic theory progresses using register- and survey data. The second part will move into the fairly new field of behavioral economics, where economists have started to understand that the traditional assumptions of *homo economicus* could not explain all behavior. Behavioral economists have adapted the use of experiments in economics, which up until the 1970's was considered a data generating method, not suited for economics. This brings me to the third and last part of this chapter, which will go through experimental economics.

2.1 Traditional economics

The term *Homo Economicus*, or economic man, has been around in economic literature for a long time. In his summary of the origin of the economic man, Persky (1995) states that the earliest naming he could find was in John Kells Ingram's A History of Political Economy (1888). The economic human is the ideal model of a human being according to many traditional economists; he acts time consistent, fully rational, and highly self-interested.

Rationality, as perceived by economists, states that individuals are able to rank his preferences in a consistent way. Preferences need to be complete, reflexive and transitive. Other assumptions state that the individual would always prefer more to less, which follows axioms of monotonicity and local non-satiation. By combining several assumptions, the *homo economicus* acts as he is maximizing utility as a consumer, and economic profit as a producer. For a formal description of a rational economic agent, see utility maximization and the theory of consumer preferences in a standard microeconomic textbook, such as Varian and Norton (1992).

As I have presented, traditional economics are based upon strict axioms and assumptions about individuals' economic behavior.

Progress of Economic theory

The engine of progress in every scientific field is how theory and empirics influence each other (Friedman, 1994). According to Carl Popper's principles of falsification, theory should be tested empirically, and if some small or major deviations from theory occurs, theory should be modified (Popper, 1954, Popper, 1959). If economics where to follow this rule of falsification, a lot of economic theory could be considered unscientific, given its extreme simplifications and strict axioms. However, Lakatos (1980) argues that a theory should never be revised in isolation, but always within its paradigm, or scientific research program, consisting of a larger set of theories and methods, which is the case for economics.

For a long time the only available economic data used to revise economic theory, was register data, and reluctantly, survey data. Economists developed and used tools within econometrics, such as multiple regression analysis, in order to mimic the *ceteris paribus* condition by introducing control variables.

However, controlling for confounding effects by the use of control variables may be seen as an imperfect substitute for doing a controlled experiment with different treatments. The following part will address behavioral economics, where psychology plays a larger role of the economic theory, and its main source of data generation, economic experiments.

2.2 Behavioral economics

By including elements from psychology, economists started to explore which implications social, cognitive and emotional factors have on the way humans make economic decisions. Economists found that individuals did not act totally self-interested, but rather to be reciprocal and care about fairness.

Contrary to what should be expected of the economic man, who would strive to get it all for himself, surveys and economic experiments show that, many people view some inequalities as fair (Almås et al., 2010). Most people would find it fair that their more productive, or higher educated co-workers have a higher wage than them, and some would argue it fair when a bigger share of public funds are allocated to projects producing the greatest benefit for the population (Almås et al., 2010).

When investigating fairness, the dictator game is often used. The dictator game is a two-player game, in which one of the two players are chosen to be the dictator, and the other is the receiver. The dictator gets a sum of money from the experimenter, which he has to decide how to divide between himself and the receiver. The receiver has to accept the offer, no matter the allocation. Traditional economic theory, predicts that a dictator with complete control should give nothing to the receiver. However, this is seldom observed experimentally. Even when the dictator is totally anonymous, the dictator gives something to the receiver in about 40% of the cases (Hoffman et al., 1996).

In one of their treatments, Cherry et al. (2002) found that when the dictator has to work for his initial endowment, he becomes more self-interested and offers less to the other player, than observed in previous experiments. While bargaining over earned wealth could have made the dictator more selfish, it could also have triggered a different norm. The dictators in the experiment could have felt that the additional surplus made belonged to them. Since they had worked for the endowment, they were entitled to a larger share.

As I will explore and present, several things affect how we make decisions, and not everything affecting our decisions will seem rational. Being humans, not econs, we are prone to several biases when we are making decisions, which will not always lead us to the "rational" and perfectly calculated expected decisions.

Individuals base their decisions on three grounds: logic, statistics or heuristics (Gigerenzer and Gaissmaier, 2011). The classical view of rationality assumes that all decisions are based on perfect information on all alternatives. However, this may be a somewhat misunderstood concept. As stated by Savage (1972), this perfect knowledge could be seen as "small world" and should be distinguished from the "big world", where perfect information is not possible. He did not suggest that individuals always have perfect information. Even when Samuelson (1937) wrote his note on how to measure utility, he stated that this was done based on several assumptions. He did not claim that this is how the real world works; he just made utility fit in a mathematical model.

During the 1970's, psychologists like Tversky and Kahneman (1973) started to explore "errors" in human reasoning. One of the explanations for these errors was heuristics. They showed that in order to make decisions easier, we develop a system of strategies, conscious or

unconscious, that allows us to ignore information, in order to make the decisions easier. Until the last decade, heuristics have been treated as bias or anomaly in the human cognition, as compared to "rational" decisions. However, now, heuristics are more accepted as a method of making decisions, rather than being viewed as a cognitive error (Gigerenzer and Gaissmaier, 2011).

In the following section, I will present some heuristics, which stems from the emerging field of behavioral economics.

Loss Aversion

The theory of loss aversion tells us that the impact of losing something affects us about twice as much as the impact of gaining something (Tversky and Kahneman, 1991). In their paper on prospect theory (Kahneman and Tversky, 1979), they questioned the reigning theory of expected utility, pointing out that it didn't pay any attention to the individuals reference point. The expected utility theory, only measured absolute value changes in wealth or welfare, rather than the final state. When evaluating absolute values of changes in wealth, you don't pay attention to the persons current assets. They argue that value should be treated as a function of two arguments, the reference point and the magnitude of change. As an example of why reference point matters, imagine trying to decide whether a bowl of tempered water is hot or cold without any other method of measurement besides using your hand. If your hand were freezing, the water would feel warm, and vice versa. In addition to include the reference point and magnitude, the value function is generally concave above the reference point, and convex below. That is, concave for gains and convex for losses. This implies that the feeling of a gain would have less impact than the feeling of a loss. A loss of $100 would have roughly about twice as big an impact as gaining $100, as shown in Figure 1. The s-shaped curve represents the value function graphically, and we can see that the perceived loss is about twice as big as the gain. For a mathematical approach to the value function, see (Thaler et al., 1997)

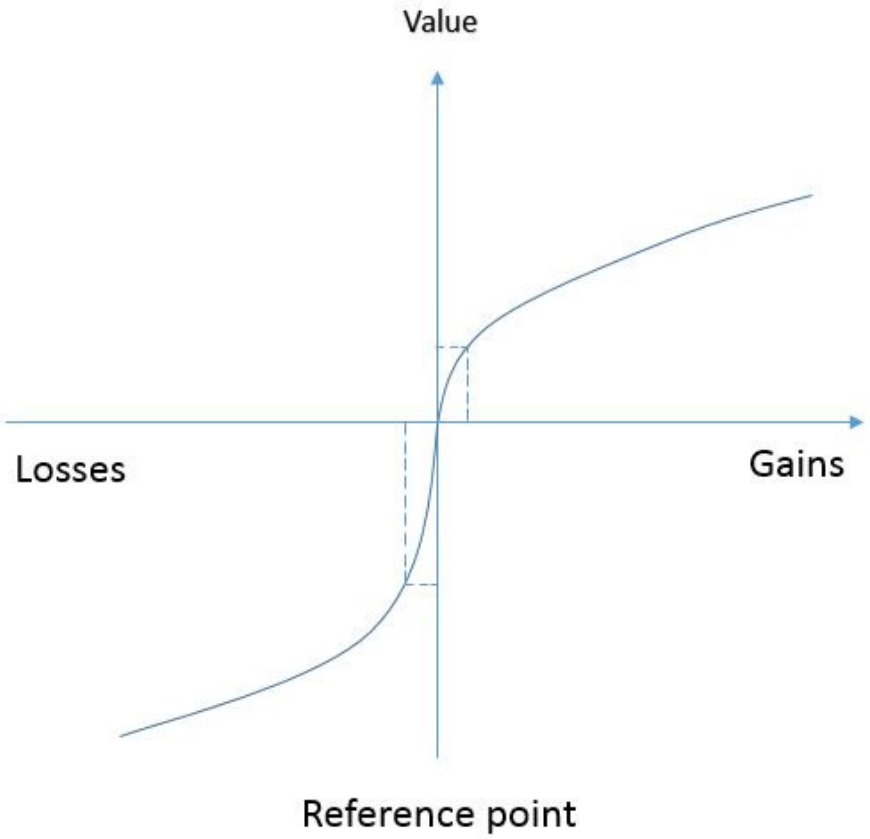


Figure 1: Loss aversion, the value function.

Status Quo Bias

Status quo is defined as the currents state of affairs, and it that a status quo bias is an implication of loss aversion. An individual facing a decision to leave status quo, is likely to stay there, because the negative effect of leaving status quo would impact the individual more than the positive effect of a possible gain of leaving the status quo (Kahneman et al., 1991). It is obvious that this would have an impact on decision-making.

Samuelson and Zeckhauser (1988) was one of the first to demonstrate this effect by conducting several decision making experiments in order to see how likely people were to stick to status quo. In the classical model of choice under certainty, individuals are able to transitively rank all options according to their preferences and choice axioms. This makes it possible to choose the highest-ranking alternative at any given time, making it the rational

choice. The individual would not be distracted by irrelevant options, and we would know when we observed the choice, that it indeed was the individuals' highest ranked alternative. One of the basic principles of rational choice is that only the values affecting ones preferences matters when choosing an alternative, the choice will not be affected by framing or in which order the options appear.

Samuelson and Zeckhauser (1988) used questionnaires in order to examine if their subjects inhibited a status quo bias. The subjects, who played a role of a decision maker, was asked to state their preferred choice among several alternatives. In the experiments, the wordings of the decision making was altered. In the first part of the questionnaire, the subjects were told that they inherited a large sum of money, and asked how to invest it. And in the second part, which was the status quo point, they were told that they inherited a stock portfolio, and was asked if they would reallocate the funds. Their main findings was that the subjects, to a large extent, had a status quo bias. The more alternatives the subjects could choose from, the stronger was their relative bias for the status quo, and if the subjects had a strong preference for a certain alternative, the bias was weaker.

It seems easier for the decision-maker to stick to the status quo, rather than use a lot of effort in order to deviate. The decision to stick to the status quo, could be driven by the fear of loss aversion by leaving, or some sort of heuristic in order to extort less effort in making the decision.

As for this thesis, the status quo state would be to give the candidate a C, as that is the expected average of the students, and easiest to make an argument for after the grading is done. In a case where the censors are in doubt whether the candidate would get the grade D or B, giving the candidate a C, could be the easiest thing to do. Hence, the status quo bias, could lead to more students getting C's when in fact, they actually deserved getting either a better, or a worse grade.

Anchoring and adjustment

Tversky and Kahneman (1974) conducted an experiment where a rigged wheel of fortune decided a number for the participants. The participants received a number (in the experiment,

the only two possible numbers was 65 and 10). When the number was given, the participants were asked if the percentage of African countries in the United Nations were more or less than the given number. First, let's see what happened when the number was 65. Most of the participants stated that, of course, it had to be less than 65 percent. The next thing they were asked, was "what is the exact percentage of African countries in the UN?". After some thought, the mean of the answers was 45. The same thing was done to the next participant, only now the wheel stopped at 10. The average participants answer to the first question was now, that there were more than 10 countries. When asked to state exactly how many, the average of the answers was 25. The participants did not actually know the answer to the questions, but somehow the obviously random number they were assigned had a positive correlation to the guessed percentage, even if the number should have treated as totally irrelevant to the question. It seemed that when the participants received a "random" number, they considered it (the anchor), and decided whether or not it was too high or too low. After the consideration, they adjusted their estimate according to the anchor given, in the direction they found appropriate. When they were asked if the percentage was over or under 65% they correctly considered it to be less, but failed to adjust down enough, and ended up with a high percentage. This was coined anchoring and adjustment.

In order to investigate if anchoring had an effect on peoples actual willingness to pay (WTP) for familiar consumer products, Ariely et al. (2003) used the last two digits of participants social security number. Fifty-five students were presented to familiar consumer products, like wine, chocolate and books with a mean price of $70. Participants were then asked if they were willing to pay a sum corresponding to the last two digits of their social security number. After they stated yes or no, they were asked what their maximum WTP would be. It was then randomly decided which of the two sums the participants had to pay, which all the participants was made aware of. The results of this experiment, was that the social security number indeed had a significant effect on the WTP. In fact, participants with above median social security number stated a WTP of about 57-107% higher than those below.

This study has been replicated by Bergman et al. (2010), which in addition to investigating how the social security number affected the WTP, also checked if the anchoring effect was related to the individuals cognitive skills. The procedure in the replication part of the study was comparable to that of Ariely et al.'s study. Immediately after the subjects had stated their

WTP and before the actual purchase, the subjects were set to conduct two cognitive abilities tests. In the first part of the experiment, their results was that the willingness to pay was indeed positively correlated to the social security number, as predicted by the anchoring effect. In the second part they found that the anchoring effect was smaller when the subject had higher cognitive skill, but that the effect was still there, even in the subjects with above median cognitive skills.

As we have seen, the anchoring effect have proven both robust and reliable through several studies, and has even been made measurable by an anchoring index (Jacowitz and Kahneman, 1995). The anchoring effect affects our judgment when making decisions; it could be that this effect also could affect the grading of the student.

The empirical methodology of behavioral economics is by and large experimental. It is often difficult to discuss one without the other and I have frequently referred to experiments in the section on behavioral economics. However, I will now turn to issues specifically linked to experimental economics.

2.3 Experimental economics

In an early version of Samuelson and Nordhaus (1985) textbook, Principles of Economics, they stated the following:

> *"Economists … cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe."* (Samuelson and Nordhaus, 1985, p.8)

In his introduction to experimental economics, Friedman (1994) writes that it is not only economics that have been claimed to be non-experimental. The same was true for physics in Aristoteles's days. Biology is another example of a field, which was considered non-experimental, since it dealt with living organisms. Now a day, both physics and biology is highly experimental.

For a long time economics was also considered to be non-experimental, but starting with the experimental work of Vernon Smith, Charles Plott and others, experiments are now commonplace. The field of experimental economics has exploded, and its methods are used within a wide array of economic theory, like industrial organization, game theory, finance and public choice (Card et al., 2011).

Experiments in economics have several advantages over the more traditional methods, such as register- and survey-data. Economists conduct experiments in order to be able to tell if effects are causal, and to observe actual behavior.

In the following part, I will address the issues in internal versus external validity, and some main differences between experimental methods used in psychology and economics, since they are both social sciences.

Internal validity

An important factor when running economic experiments is that it has to have high internal validity. First, the researcher should construct and present the theoretic model. Second, predictions are generated deductively, and third, the experiment is run to test the predictions (Croson, 2005). In order to obtain high internal validity, the researcher has to construct a laboratory situation, which are in line with the theoretical assumption of the theory. If the researcher for instance, has an assumption of a one shot interaction, the experiment cannot be run using repeated interactions, and vice-versa.

By controlling all factors of the experiment, using proper control-groups and randomization, the experimenter have the opportunity to see which causal mechanism really gives the results. If the experiment is run properly, with an open protocol, it would be easy for others to replicate it, ensuring that the result is robust.

A problem with experiments is the possibility that a measured effect is an artefact of the experiment. This means that the result is biased by the experiment itself. Another problem is demand-effects, that the subjects acts in a way they think the experimenter expects (Zizzo, 2010). This is possible to counteract by paying close attention to the context given and

ensuring that the subjects understand how the experiment works. More on this follows, in the *Rules* part of this section.

External validity

Even if an experiment is run by proper standards, and it produces robust replicable results, it has no guarantee of being important for real-world problems.

One of the problems with obtaining external validity is the subject pool. If the subjects are not representative of the population, the results obtained in the experiment is difficult to generalize. Cappelen et al. (2011) examined if there were any differences in the results of student participants and representative participants in the case of social preferences. By having one treatment with subjects from a representative selection of the population and one treatment with students, they clearly demonstrated that the behavior of the students in the experiment differed fundamentally from the behavior of the representative group. This indicates that by conducting experiments on students, we may not be able to generalize results to the real world. Their experiment specifically examined social preferences, and their result may not generalize to other aspects of social sciences, but it gives an indication that it might.

Being representative of the population is not the only problem; subjects can also be under-experienced. If for instance you where to conduct an experiment with the goal to see whether fishermen would overharvest a resource stock (common-pool resource dilemma), you would arguably have more external validity if you did the experiment on experienced fishermen, rather than students.

While bringing your experiment out in the field might give you higher external validity, you may encounter other difficulties, such as under-incentivized subjects, or if the problems your subjects face in the experiments are too alien or non-naturalistic (Uuskartano, 2013). Some of these issues will be addressed shortly, in the *Rules* part.

Rules

There are several "rules" as how to run an experiment, which differs between Psychology, Experimental economics and Behavioral economics. A very good review of this is done by Croson (2005). In this thesis, I will go through the three, which I feel is the most important; incentives, context and deception.

Incentives

Economic theory describes and predicts what actions individuals will take when offered some economic payoff. As an example borrowed from Croson (2005), we can use a public good game. When deciding how much to contribute to a public good game, individuals have some private cost. Since the cost of contributing is larger than the cost of not contributing, and the received benefit from contributing is less than the received benefit when not contributing, the prediction is that individuals will not contribute.

As experimental economists are interested in what individuals actually do, and not what they say they will do, it is important to pay participants according to their decisions. In psychological experiments, it is often stated that money is not the only motivator, whereas in experimental economics monetary incentives in experiments are required. Since cash is usually perceived as non-satiable, the participants in economic experiments are often paid in cash after the experiment is ended. Friedman (1994) suggests using subjects with low opportunity costs, such as undergraduate students, since paying more than the subjects average opportunity cost, should promote monotonicity and salience.

In a meta study on how performance based financial incentives affect experimental results, Camerer and Hogarth (1999) reviewed 74 experiments with no, low or high performance based financial incentives. They found that incentives improved performance in easy effort responsive tasks, such as judgment, prediction and problem solving, but if the tasks become too difficult, it sometimes had negative effects. When they examined results from auctions, games, and risky choices, they found that incentives did not affect the mean outcome, but it

often reduced the variance. According to Croson (2005), there is a general agreement that contingent payment reduce the variance of responses in experiments.

One downside of using contingent payment in experiments is that it can get expensive fast. In order to keep experiments within their budgets, researchers can manipulate the payment scheme (Croson, 2005). For instance, if participants play a repeated game of 20 rounds, the researcher can explain the participants *ahead* of the experiment that only five of the rounds will be paid in real money. When the experiment has ended, five of the rounds will be chosen at random, and participants will be paid according to performance. This way, if the experiment had 20 rounds, the researchers only have to pay 25% of the cost, as opposed to if all rounds were paid. Since the participants did not know which rounds would be paid, they would perform as if they were paid in each round. This method can also be used in experiments with a large number of participants. The researchers could for instance randomly determine one of the treatment groups, which will receive payment, after the experiment has ended.

In sum, incentives are very important to economists when designing experiments. Incentives has been shown to reduce variance around the predicted outcome, bring auction bids closer to optimality and reduce preference reversals (Croson, 2005).

Context

While context is often rich in psychological experiments, economic experiments are often context free. Croson (2005) provides three reasons why economists prefer their experiments to be context free. First, the theories tested are supposed to apply generally, the predicted outcome should not vary with context, as long as the payoffs are the same. Her second reason is that that context often leads to additional variance in the data. While the additional variance might not change the outcomes directly, it can reduce the likelihood of detecting statistical differences between treatments. And the third an most important reason according to (Croson, 2005), is that context may add bias or demand effects. If a problem is framed in such a way that the subjects understands that the experimenter is looking for specific treatment effect, they may act in that way in order to “please” the experimenter, which would lead to a biased result.

As pointed out by Loewenstein (1999), if the experimenter would construct a truly *"context free"* environment, where everything is sterile, that too would be a context, maybe more alien than most other environments. Loewenstein goes on explaining that if the instructions become too unfamiliar and context free, *"Subjects may seem like zero intelligence agents when they are placed in the unfamiliar and abstract context of an experiment, even if they function quite adequately in familiar settings"* (Loewenstein, 1999, p.30)

The context of the experiment should be as simple as possible and still making you able to address your research questions. By having little context, you can avoid ambiguity when interpreting you results, but you should make sure you are clear enough in you instructions so that your subjects understand how to complete the experiment (Friedman, 1994).

Deception

One of the strictest rules of experimental economics is that deception is prohibited (Croson, 2005).

Croson (2005) argues that an experiment in economics require full information on the purpose of the experiment, whereas Friedman (1994) suggests that one should not inform the participants of your experimental goals. There might not be a strict rule as to when to keep the information on the purpose of the experiment private; rather it should be private when appropriate. If you are investigating racial discrimination, there is no way you can tell your participants what you are looking for, without having the participants adjust accordingly.

As mentioned in the incentives part, if researchers are to choose one of the treatments at random which is paid, participants has to get this information before the start of the experiment. If participants are told that only one treatment is paid *after* the experiment, this could lead participants to act different and distrust the experimenters in future experiments. Friedman points out that *"… experimental economists require complete credibility because salience and dominance are lost if subjects doubt the announced relation between actions and rewards, or if subjects hedge against possible tricks.* (Friedman, 1994, p. 17)*"*

One (in)famous experiment from social psychology using heavy deception is the Milgram Experiment (Milgram, 1963). The experiment deceived subjects both on the purpose of the

experiment, and that the subjects had opposing subjects. Such experiments relying on heavy deception may cause a pollution to the subject pools, where subjects have heard about, or experienced earlier experiments using deception, causing subjects to expect it. This would obviously affect subjects' decisions in experiments, making it harder for other researchers to conduct experiments with meaningful results. This is one of the main reasons for experimental economists not to allow deception. Croson (2005) claims that economic journals will not publish papers with experiments based on deception, where as in psychology deception is still commonplace.

"The Rules"	Psychology	Experimental Economics	Behavioral Economics
Deception	Allowed, if justified	Prohibited	Almost always prohibited
Incentives	Rare; Money isn't the only motivator	Required	Generally used
Context	Often rich	Stripped away	Sometimes studied

Table 1: Rules in experimental methodology

As we have seen, many considerations has to be taken in order to carry out a successful experiment. The researcher has to weigh internal and external validity against each other, as well as taking great care in designing payment schemes, recruiting subjects and deciding how much context to include. It is important to understand that theory and both laboratory and field research is supposed to complement each other, and they all count when drawing conclusions about results. As Harrison et al. (2011) puts it, *"Any data generated by an experiment needs to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods and expected applications"* (Harrison et al., 2011, p.1)

The following part contains a brief explanation about the three most common types of experiments: laboratory, natural and artefactual field experiments.

2.3.1 Laboratory experiments

Laboratory experiments is exactly what it sounds like, an experiment run in a lab. By running an experiment in a lab, researchers are able to keep the environment and context as sterile as they like. When conducting a proper laboratory experiment, researchers are able to control all relevant factors. While experiments differ in goal and scale, the basic idea is to create a small environment in which researchers have the opportunity to manipulate all relevant factors, in such a way that they are able to determine what causes the effects.

By randomizing subjects into a control group (without any treatment) and experimental groups (with treatments), and comparing the results, the researchers are able to determine if the treatment caused the measured effect, or not. While an experiment cannot *prove* a hypothesis, it can provide evidence to support it, or vice versa.

An example of a laboratory experiment could be when examining placebo effects. Researchers could give one group a sugar-pill with no anesthetic effect (placebo group) one group a pill with anesthetic effect (treatment group) and one group no pill (control group). After waiting the appropriate time in order for the anesthetics to work, the researchers can inducing pain on the subjects, for instance by small electrical shocks, and having all subjects rate the pain on a scale from one to twelve. After collecting data from the appropriate number of subjects, the researchers can compare the results from all groups, and determine how big the effect the placebo had on the subjects. Let us say the placebo group reported significantly lower pain than the group receiving no pill at all. In this case, the researcher are able to say that there is a placebo effect present, given that there were no other difference between groups. If the researchers for some reason did not include the control group, and the subjects reported lower pain in the placebo group, they would not be able to tell how big the placebo effect was; only that it did not work, compared to the anesthetic.

2.3.2 Artefactual field experiments

The main difference between a laboratory experiment and an artefactual field experiment (AFE) is the subject pool. Where laboratory experiments usually recruits standard subjects (in most cases students), AFE takes the laboratory experiment out in the field with non-standard subjects (Harrison and List, 2004). As previously mentioned, if you wanted to examine how fishermen act in a common-pool resource dilemma, it is hard to justify using students as subjects. It would be easy to argue that using non-standard subjects could ensure better external validity.

2.3.3 Natural field experiments

A natural field experiments means that you find natural occurring data, that is, you do not generate the data in a lab.

As previously mentioned, the subjects in an experiment can act according to their beliefs about how the experimenter expects them to behave, which in turn can result in a systematic bias in the results (Zizzo, 2010). When conducting a field experiment, this cannot happen, since subject *do not know* that they participate in an experiment. When conducting laboratory experiments, there is a possibility that you attract subjects who feel they have the most to gain from participating. By conducting a field experiment, the subjects do not know they participate; hence, you eliminate the problem. From the example of a pain study, you could attract people who tolerate pain rather good, so they would assume participating in the experiment is "easy money". While this would still give you valid causal results, the fact that you lack subjects from the proportion of the population who would not participate in such an experiment, makes it hard to generalize to the population (List, 2011, Levitt and List, 2007).

On the other hand, the fact that subjects do not know they participate makes it hard for them to give a voluntary consent. The fundamentals of voluntary consent is that the subjects are physically able to give consent, are not coerced into giving it and that they understand risks and rewards from participating. One could easily argue that voluntary consent is essential when dealing with medical experiments, since it could involve great risks, but in the case of economic experiments, the case is not as clear cut (List, 2011).

There are various reasons why people volunteer for experiments; this could be self-regarding preferences such as the monetary payoff, or prosocial preferences such as contributing to research. In order to investigate which factors contribute the most in the self-selection process of economic experiments, Abeler and Nosenzo (2014) designed a field experiment, in which first year student-subjects were randomly assigned to one of three treatments. The students received one of three email requests to join an experimental subject-pool, depending on which treatment they were in. In the first treatment, they emphasized the potential monetary reward for participating in an economic experiment (money only), in the second treatment they only emphasized the importance of helping research (appeal only), and in the last email they mentioned both (money and appeal). They found that the sign-up rate dropped by about two-thirds when the monetary payments were not mentioned. The percentage of students contacted, who signed up from the "money only" and "money and appeal" treatment, were 14.6% and 13.8%, but these groups were not statistically different from each other. In the appeal only treatment, however, the signup rate was only 5%, which was significantly lower than in the other treatments. The results found strongly suggests that the monetary rewards for participating in economic lab experiments is the main reason for participating.

As we have seen, there are pros and cons for doing a field experiment as well. While the data is naturally occurring, providing higher external validity, it is less controlled than in a laboratory and the results may be affected by factors you do not control. You could also have problems with sufficient randomization and it is hard to get voluntary consent from your subjects.

To sum it up, there are several types of experiments which can be used to generate data, some are more controlled, and some using naturally occurring data. While more control gives higher internal validity, natural occurring data provides higher external validity.

Controlled data	**Naturally-Occurring Data**
Laboratory experiment	Natural field experiment
Artificial Lab experiment	Quasi experiment

Table 2: Experimental data

The two studies I will present providing evidence for the lunch-effect are both based on quasi-field experiments. The data is generated without subjects being randomly assigned to treatment groups. The subjects in Danziger et al. (2011a) is the judges, with the treatments being before and after food-breaks. The same is the case on clinical decisions on the use of antibiotics. The current research is also based on data collected where the censors are the subjects, and the treatment is whether they have had lunch or not.

3. Background for the experiment

In the following parts of my paper, I will first go through some sources of grading biases as listed by Fleming (1999), before having a look at the some literature concerning mental depletion and it's critique. The last part of this chapter concerns "the lunch effect". It consists of a thorough presentation of Danziger et al. (2011a)'s paper a replication by Linder et al. (2014), and critique to Danziger et al. (Weinshall-Margel and Shapard, 2011).

3.1 Grading biases

One can hardly argue against the fact that grades should be given based on the students' performance. However, research suggests that several biases exits, which may affect the grade given. A bias in this context will be defined as *any factor contributing to the grade, not related to the performance of the student.*

The implication of giving out biased grades could, as I have mentioned in the introduction, be that students chooses education which is not optimal, the loss of student loan, and it could also be a part in the explanation of the increasing wage gap between men and women.

The following section will go through some of the biases presented by Fleming (1999).

3.1.1 Gender bias

Bernard (1979) examined whether a teachers evaluation of a student is solely based on the perception of the students' performance, or if the evaluation is influenced by stereotypical sex-roles of the students. The subjects in this study were two hundred and forty teachers, half of which were female and half of which were male. The teachers evaluated a male or a female student portraying masculine or feminine behavior having a major interest in English or physics. In order to do this, he constructed eight different seven-page booklets. The first two pages in the booklet was identical, it informed the subjects that the study was looking to investigate some variable that might affect the teachers evaluations of students, and to see if certain characteristics of students can be accurately estimated from description. The booklet further asked the subjects to record a number of demographic variables about themselves, as age, number of years teaching etc. The third page, gave the same brief description of a student. The twist here is that the student is either named John or Jane Stevens; this is the first experimental manipulation of the experiment. The description portrayed John or Jane as a 17-year-old student, liked by friends, perceived as motivated and generally interested in school by her/his teachers.

After the general description of the student, a second experimental manipulation occurs. This part gave the student a stereotypical sex-role behavior; the student is given either a masculine description or a feminine description, which is crossed with sex and major field of interest (English or physics). The subjects are then first asked to rate on a scale from 1 to 7 how well certain characteristic fits John (Jane). The characteristics are intelligence, warmth, masculinity, hard work, independence, logic, concern for others. After this they are asked how much they would like to have this student in their class, how well the students choice of course fits them and finally they are asked if they think the student will have difficulties doing their chosen subjects at a university. The fifth page was a short answer of either 228 or 301 words, where the "student" answered a physics or an English question, both answers where considered above average in quality. The sixth and seventh page contained the same 10 questions as before rated on a 1 to 7 scale, and then the subjects were instructed to evaluate the student on several dimensions.

Bernard (1979)'s main results revealed that the student sex-role behavior affected both the subjects' perception (teachers'), as well as their evaluation of the students' written performance. In fact, the results indicated that teachers expected higher intelligence, independence, logic and academic success from masculine sex role behaviors whereas feminine sex role behavior was associated with warmth and concern. Students who was associated with masculine sex role behavior was assessed as superior to students associated with feminine sex role behaviors, in fact, feminine sex role behavior was viewed as negative for students studying physics. These findings could provide some evidence that students sex role behavior, masculine or feminine, could affect the way censors evaluate the students written work.

By examining data from a natural experiment Lavy (2008) set to explore the existence of gender stereotyping and discrimination by Israeli high-school teachers. Israeli high-school students can enroll into two different tracks of education. They can choose either the academic track, which lead to a matriculation certificate, or the vocational track leading to a high-school diploma. The final matriculation grade in a given course is the mean of two tests; the first is graded by the school and the other is graded by an external independent agency. The external censors only have a student ID number, assuring anonymity of the student, while the internal censor is the students own teacher, which is not anonymous. The tests take place at the students' school, and in regular classes, which should control for externalities, such as the possibility for higher anxiety levels. This makes good basis for a natural field experiment, with two treatment groups; one which are blind (external censor do not know the identity of the student) and the other is non-blind (teacher is the internal censor). In the blind group, censors have no way of being affected by gender stereotype biases whereas the non-blind group may be affected by such biases. The data in this field experiment was from the school year 2000-2002, it included all matriculation grades (nine subjects) for each student during their high school (grades 10-12), and student characteristics such as gender, parents' schooling, family size, etc.

By using the difference between the boys' and girls' gaps between the blind and non-blind test scores, Lavy (2008) was able to measure the potential gender bias. Lavy found discrimination against gender in the data. Contrary to his expectations that this discrimination would be against females, he found that there was discrimination against males in all nine

subjects examined. He concludes that this is a result of discrimination resulting from teachers behaviors. This main finding, that males seem to be discriminated against, is supported by research by Lindahl (2007a). She finds, by comparing non-blind (school leaving grades) against compulsory national test scores, that females are given better grades in the national test, than males. This result was the same for all three subjects studied by Lindahl.

As pointed out by Hinnerich et al. (2011) a limitation to both Lavy's and Lindahl's study, is that the non-blind and blind test is not exactly the same, and the fact that both students and teachers know that one of the tests are graded locally, may affect the test scores.

Hinnerich et al. (2011) drew a random sample of 2880 students from 100 schools in Sweden where the students had done a national test in Swedish graded by their teacher. They had all tests rewritten in a word processor and removed all teachers notes and student identification. After this, they hired 42 experienced teachers, which was familiar with grading national tests, to re-grade the tests. By doing this, they now had a non-blind and a blind group, with exactly the same test. Due to various reasons explained in the paper, Hinnerich et al. (2011) only obtained 1712 observations. They found that females on average obtained better grades than males, in fact the average non-blind test score was 15% lower for males than females. The data on their blind tests shows that the grades decreased by 13%, but the decrease was almost identical between males and females, hence, they found no evidence of discrimination. As they point out in their concluding remarks, the difference between their result and the result of Lavy (2008) can be due to the fact that the Swedish study only examined results from one subject, whereas Lavy examined nine, and that discrimination could exist in other subjects in Sweden. It could also be that discrimination occurs in Israel, but not in Sweden, or that the difference in scores in Lavys study is due to other factors than discrimination.

3.1.2 Halo effect

If the person grading the student have knowledge of the students' previous achievements, it may influence the grading.

In an experiment Nisbett and Wilson (1977) instructed two groups of students to watch a recorded interview of a psychology instructor. The instructor, who was a French-speaking

Belgian who spoke English with an accent, appeared different in the two tapes. In the tape shown in one group, the teacher appeared likable, respectful and flexible, whereas in the other he appeared unlikable, cold and distrustful. After viewing the interview, the students were asked to evaluate the instructor on several issues, including likability, attractiveness of his physical appearance, his mannerism and accent. Some of the student where asked if the instructors likability had affected their rating, in order to explore subject awareness. Other students were asked the quite opposite, if their rating of the instructor had affected their rating of his likability. Nisbett and Wilson (1977) expected that the students evaluating the instructor would rate him higher on physical attractiveness, mannerism and have a more attractive accent, when he appeared likable as opposed to when he was unlikable. They also expected that the students would be affected by the appearance of the instructor when doing their ratings, without knowing so, and would deny that they were affected afterwards.

The results was that the students who were shown the warm instructor, rated his attributes as appealing, and the students shown the cold instructor rated his attributes as non-appealing. More interestingly, when asked if their like or dislike of the instructor had affected their rating of his attributes, they reported that it had no effect on them. Students who saw the unlikable instructor, actually believed that the direction of influence was opposite of the true direction. They reported that their dislike of the instructor did not affect their rating of his attributes, but that when they found his attributes low, it affected their likeness of him.

This experiment shows that we can unknowingly, be affected by our initial favorable impression of a person when we are to subjectively rate their performance, causing us to give them higher evaluations later. It is however not always the case that it is unknowingly, the grader could also be aware of his bias. For a more anecdotal paper on self-awareness of biases in grading, see Malouff (2008).

To assure anonymity students often receive a candidate number. Professors could still be able to recognize the student's gender through their handwriting and/or writing style (ie. Sentence structure). In the case of an oral exam, the student are not anonymous.

3.1.3 Other effects

The performance of the previous candidate may influence the grading of the next candidate. The effect is may be better explained by an experiment done by Kenrick and Gutierres (1980). In their experiment, they went into dorm-rooms where students was watching the, at the time, hit TV-show Charlie's Angels. They asked the students to rate the attractiveness of a women photographed on a scale from 1-7. Their experiment revealed that, when men was asked to rate the attractiveness of an average looking woman, right after watching three attractive women, they rated her significantly less attractive than the control group, who had not been watching Charlie's Angels. Although, the study may have had its flaws, such as no randomization into treatments, and that maybe people watching Charlie's Angels have a more negative view again women, the results may provide evidence towards a contrast effect. In the case of grading, the censors may be viewing an average exam as less than average, when grading it right after an outstanding exam.

Ethnic biases may also exist. By examining certificates by teachers and results from compulsory national Swedish tests, Lindahl (2007b) found that when the proportion of ethnical minority teachers increased, students with ethnic minority background on average obtained better grades. This points to that the ethnicity of the candidate could influence the censors. Another possibility, pointed out by Lindahl, is that the teacher serve as a role model for the student. Having an ethnic minority teacher influences the ethnic minority students to do better.

As we have seen, as humans are not computers, able to assess all given information in a truly objective way, several things may affect their decisions. In the case of grading, things such as gender, ethnicity, halo-effects and contrast-effects may contribute to the grades given by the censor. The following part will have a look at ego-depletion, which may also influence the way we make our decisions.

3.2 Mental Depletion

Research suggest that people have a limited resources available for self-control, and that making decision by an individual will eventually deplete this resource (Muraven and Baumeister, 2000b, Baumeister et al., 1998). In Muraven and Baumeister's paper they argue that self-control resembles a muscle, which will tire when its available energy reserves are depleted by self-regulatory exertions, making it vulnerable to fail after strenuous use. Kouchaki and Smith (2013) found using four experiments, that people where more eligible to cheat or act in immoral behavior later in the day rather than earlier. In the four experiments conducted, the subjects were set to do relatively easy tasks, like reporting whether or not there were more dots on the left or right side of a screen or solving matrices where half of them had a solution, and the other half did not have a solution. The subjects where paid ten times as much if the right side of the screen had more dots than the left, and in the matrix problem they were paid according to how many matrixes they solved. Both cases gave the subjects the opportunity, and monetary incentives to cheat. The experiments were run both in the morning and the afternoon. All of the four experiments showed that the subjects where more eligible to cheat/act immorally in the afternoon experiments. They argue that everything we do from the moment we get out of bed, every decision we have to make, no matter how trivial, it will drain our mental capacity leading us to be mentally exhausted which in turn can decrease moral awareness and lower our self-control in the afternoon.

If this is the case, that even the most trivial of things can drain our mental capacity, then it would be plausible to think that censors evaluating students the whole day, will also be affected by mental depletion.

There have, however, been done research in order to see if there are any ways of restoring this mental capacity. Tice et al. (2007) found by conducting several experiments where the subjects were set to do ego-depleting tasks, that by letting the subjects watch amusing comedy videos or giving them a surprise gift, counteracted the effect of ego depletion. As a further research in how to replenish the mental capacity, Tyler and Burns (2008) tested if relaxation and time could have a positive restoring effect. In the time experiment, they found, in a two stage test where the subjects were set to do self-regulatory tasks, that giving the subjects a 10 minutes break between the tests, greatly enhanced their performance in the second test, compared with those given a 1- or 3-minutes rest. In this experiment, the subjects were given

a task to fill out some forms while having the break, whereas in the second experiment the break was more relaxing. They found that the 3 minutes of relaxation was also reducing the typical depletion effects. Gailliot et al. (2007) found support for their hypothesis that acts of self-control led blood glucose levels to drop below the optimal level, and thereby reducing the control over thought and behavior.

In short, the bad news is that the mental capacity seems to resemble a muscle, which can be depleted. This may causing decision-makers to do the "easy" decisions, and sticking to the status-quo. The good news is that measures can be taken to counteract this mental fatigue, like watching amusing videos and relaxing. While watching videos may not be a good solution in the case of examinations, a longer, more relaxing break between students might have a positive effect.

It is important to mention that not all researchers agree on self-control as being a real limited resource. Some researchers find no effect of sugar ingestion to replenish the mental capacity, and others argue that it is solely up to each individuals' belief about mental depletion if it exists or not (Job et al., 2013).

3.2.1 Critique to mental depletion

A recent experiment by Lange and Eggert (2014) was designed to replicate the reported counteracting effect of sugar consumption by Gailliot and Baumeister (2007). The subjects was 70 undergraduate students (62 female), who were compensated with course credits. The subjects were instructed not to eat 1.5 hours before the experiment. Upon arrival in the lab, the subjects where assigned to either the experimental or control treatment. Subjects' initial blood glucose level was measured (T1) before they were set to do several selective attention and delayed discounting tasks. The subjects was then given either a sugar or non-sugar beverage, according to their treatment (experimental or control). After consumption, the subjects was set to rate how much they enjoyed the beverage, current state of hunger and exhaustion at an 11-point Likert scale, and to fill out some standard questionnaires which took 10-15 minutes, in order to let the sugar from the beverage be metabolized. After this, the subjects' blood sugar level was measured again (T2), before the subjects did more delayed discounting tasks.

Comparing the blood glucose levels at T2 revealed that it was significantly higher for the experimental treatment than the control, indicating that the experimental manipulation was successful. Lange and Eggert (2014) also found that the results of the second test differed between the experimental and control group, however, they found no difference in the performance between the two groups, when accounting for the pre-treatment test. The mean blood glucose level of the control treatment did not decrease between T1 and T2. This suggests that performing self-control tasks do not lover blood glucose levels, which they point out is in order with Kurzban (2009) who demonstrated that it was unlikely that blood glucose levels would decrease when doing self-control tasks.

Since the subjects participated for course credits, and was not paid by performance, It would be interesting to see the experiment repeated. The subjects had no incentives to perform their best, which might make them indifferent to their performance. If they were paid on performance, the tasks might have been more mentally draining, since the subjects would want to earn as much as possible. This could have resulted in a difference between treatments.

In order to test their hypothesis that the effect of glucose is dependent on the persons beliefs about willpower, Job et al. (2013) conducted several experiments. Eighty-seven subjects was recruited in their first experiment, and as in previous experiments in this literature, Job et al. had subjects fast for two hours before the experiment. Upon arrival, the subjects filled out six forms about the subjects beliefs and theories of the effect of mental exertion according to a previous study by Job et al. (2010). Subjects was given statements such as, “After a strenuous mental activity, your energy is depleted and you must rest to get it refueled again” (limited-resource theory) and “Your mental stamina refuels itself; even after strenuous use mental exertion you can continue doing more of it” (nonlimited-resource theory). These statements were rated by the subjects on a 6-point scale where 1 = strongly agree and 6 = strongly disagree.

After the completion of this first task, subjects were given a lemonade either containing sugar or artificial sweetener, according to their treatment, they then allowed ten minute to pass in order to let the sugar be metabolized, before the key dependent measure. During this ten-minute “break”, subjects were asked to cross out every “e” in a text, which is considered a demanding self-control task. After the subjects had completed this ten-minute task, they were

set to do a computerized Stroop-test, which is a commonly used task to measure self-control. The Stroop-test is to determine the color of the font of a word appearing on a computer screen as quick as you can. For instance, the word could be “Red” and the font color could green, in which case, the correct answer would be green (Stroop, 1935). Forty-eight of the words would be in the color associated with the word, and 48 of the words would be in a color not associated with the word.

The results revealed that subjects who displayed limited-resource theory (that mental energy was easily depleted, and could be refueled again), performed better in the sugar treatment than in the non-sugar treatment. More interestingly, for subjects displaying nonlimited-resource theory there was no difference between treatments. Job et al. (2013) argues that this demonstrates that peoples beliefs about ego depletion, determines whether glucose has an effect on willpower.

In the other experiments in their paper, Job et al. led their subjects to endorse either the belief of limited- or nonlimited-resource theory. They were then set to do a demanding (ego-depleting) task before they did a self-control task, such as the Stroop-test. The results obtained from these experiments were in line with the result of the first experiment, where they found that ego-depletion only applied to those led to endorse the belief that willpower was a limited source. For the participants led to think of willpower as a nonlimited-resource, a demanding task did not reduce the performance in the subsequent task.

Job et al. (2013) further emphasizes that these results are in line with the result of previous research done by Clarkson et al. (2010) who found that subjects who perceived themselves as ego-depleted performed worse than subjects who did not perceive themselves as depleted.

As we have seen, there are several opinions about whether self-control and willpower is a limited resource. As demonstrated by Job et al. (2013), willpower is a limited resource, only to the people believing willpower is a limited-resource. When the individual believe willpower is non-limited, Job et al. (2013) is unable to show any effect of depletion. If willpower resembled a glucose driven muscle as argued by some (Muraven and Baumeister, 2000a, Baumeister et al., 1998, Gailliot et al., 2007), it is interesting to see that experiments fail to deplete blood glucose levels as done by Kurzban (2009).

3.3 The Lunch-effect

By examining sequential judicial decisions by experienced judges, Danziger et al. (2011a) wanted to explore whether they could find any exogenous factors affecting the judges' rulings. They examined 1112 judicial rulings from parole hearings, collected over 50 days in a 10-month period. The data stemmed from two parole boards, serving four major prisons in Israel. Each parole board consisted of one judge, as well as a criminologist and a social worker, whose task is to advice the judge. In total, the data collected contained about 40% of the parole requests in Israel, and for each of the 50 days, the average percentage of parole decisions was 78.2%. Other decisions made was such as changing the terms of parole, and requests of prison reallocation. By large, we can state that the majority of cases was parole decisions. The researchers also recorded the judges' two daily meals, late morning snack and lunch, which resulted in three distinct decision sessions.

According to the theory of mental depletion, the researchers hypothesized that since the judges are making sequential decisions throughout the day, they would be more likely to make the "easy" or "safe" choice, e.g. sticking to the status quo, which in this case was denying the prisoner parole.

As pointed out earlier in this thesis, research suggests that several measures may be done in order to restore mental depletion. One of these measures is taking a short break and restoring the body's glucose levels. Since the researchers recorded the judges' food breaks, Danziger et al. (2011a) were able to examine if the breaks affected the decisions. The morning snack was served after an average of 7.8 cases and the lunch was served after another 11.4 cases.

The decisions in the case of parole decisions are "accept request" or "reject request". This makes the data collected ideal for a logistic regression, with the ruling as the dependent variable and several dummy variables, such as ordinal position in each session, which session the case appeared etc., as explanatory variables. The results obtained from the analysis was quite striking, Danziger et al. (2011a) reports that the likelihood of accepted parole started at 65% in each session, before steadily decreasing till 0% at the end of each session. As a robustness test of the result, they also did an analysis, which measured cumulative minutes elapsed in each session, as a proxy for mental fatigue among the judges. The results was

similar to the first result, giving support to their hypothesis that as time passes within a session, the probability of a favorable ruling decreased. They also found that deciding on a favorable ruling took significantly longer than an unfavorable, and that the written verdict of a favorable ruling was longer than an unfavorable. These two facts are used to argue that giving an unfavorable ruling would be the easiest choice for the judges.

A recent study by Linder et al. (2014) shows that, as the day proceeds, clinicians prescribes unnecessary antibiotics for acute respiratory infections (ARI). In this case, Linder et al. argues that prescribing antibiotics is considered the "easy" or "safe" option for the clinician. The data used for this study consisted of 21867 ARI visits by adults from May 1 2011 to September 30 2012. The clinicians doing the prescriptions had two clinical sessions each day, where the first session was from 08:00 until 11:00, followed by a lunch break, and continued from 13:00 to the end of the day 16:00. The researchers found that the likelihood of prescribing antibiotics increased throughout the two clinical sessions held each day. They state that this is in line with Danziger et al. (2011a), who argues that the "easier" or "safer" option is to refuse parole.

Both of these studies provides evidence toward an existence of a lunch effect. When decision makers do sequential decision throughout the day, decision makers tend to make the "easiest" decisions, sticking to the status quo, when nearing a food break. The provided explanation for this is that each decision uses some mental resource, and after enough decisions, the resource is spent, and mental fatigue kicks in. While Danziger et al.'s results, shows a really extreme effect, by going from 65% to 0% within each session, Linder et al. (2014) finds a more moderate effect. Whether the increase in favorable rulings, or a lower chance of prescribing antibiotics, are due to the digestion of food or merely taking a break, we do not know. Future research should examine which effects dominates. They also do not have a way of measuring the judges or clinicians' mental resources and do not know how it differs over time, but the results suggests that some exogenous factors contribute to the rulings.

3.3.1 Critique to Danziger et al.

As a critique to the paper, Weinshall-Margel and Shapard (2011) pointed out some overlooked factors in the original paper. They argue that in order for Danziger et al.'s results to be valid, the order of the cases need to be random, or at least exogenous to the timing of

meal breaks. They found that the judges tried to process all cases from one prison, before having a break and moving on to the next prison. The normal procedure here, was that the unrepresented prisoners' cases was the last in each session. It is reasonable to believe that an unrepresented prisoner is less likely to be granted parole, as they do not have an attorney present to trial their case. Weinshall-Margel and Shapard (2011)'s result using the same data as Danziger, found that the unrepresented prisoners accounted for about a third of the prisoners, and only had a 15% chance of getting a parole while prisoners with legal representation had a 35% chance of parole. They argue that this indicates that the non-random sample was the real explanation of the declining rate of success rate, rather than the lunch effect. In addition, they argued that it is normal for some prisoners to be represented by the same attorney, and that in those cases; the order of the prisoners might not be random. The attorney may have represented their "favorites" first, the ones with the highest chance of success from the attorneys perspective, and the ones least likely to get a parole last. They also pointed out that Danziger's study did not include data on prisoners' in-prison behavior.

This critiques to the original paper was answered by Danziger et al. (2011b) where they rejected that the critique was valid. They included representation as an explanatory variable and was unable to reject their original results, leaving food break to remain a robust indicator of parole. The new regression analysis showed that the legal representation was positively correlated with parole, without being a significant predictor in all the models. Additional interviews of prison personnel, revealed that the order of representation, was done on a "first come, first serve" basis, precluding any ordering by prison. The including of representation actually significantly improved the fit of the model. As for the comment on attorneys having several clients, their answer was that the number of observations was so low, that it would be mathematically impossible to calculate spikes in favorable decisions.

Although the critique made by Weinshall-Margel and Shapard (2011) seems arguable reasonable, when accounted for in the new analysis by Danziger et al. (2011b), not only did it replicating the original study, it made it even more robust and made the data fit the model better.

We are now turning to the experimental part of this thesis, where I will be examining the data collected from the oral examinations in order to see whether a "lunch-effect" exists in such a setting.

4. Data and analysis

It will be really interesting to see whether the results of Danziger et al. (2011a) and Linder et al. (2014) are replicable in another settings, such as the exam results from the University of Tromsø. The candidates in my study are called in alphabetically, which can be treated as random as the candidates name can have no effect on the grading. In addition, all candidates from one group have their examination the same day.

All students enrolled in a bachelor-program at the University of Tromsø is obligated to take a course in philosophy, and have a choice of either doing a written 6 hour school exam "distriktsvarianten", or a 20 minute oral examination "Tromsøvarianten", both which account 100% of the final grade. In my thesis, I have only collected data from "Tromsøvarianten" which is the oral examination.

Upon registering for the course, each student is assigned to a study group, according to their university program. These groups follow the student throughout the semester, and they have the same seminars, with the same teacher who also is their internal censor at the exam. It is required of the students that they attend at least 75% of the seminars in order to be able to take the exam, which hopefully leaves the students well prepared. The group sizes vary from seven to 22 students, with a mean of 16.04. During the examination day, each group is tested in the same day. Students are called in alphabetically; starting from 0800 continuing until 11:50, where the censors have a scheduled lunch, before the examinations start again 12:50 and continuing until all students in the given group has had their exam.

All data from the examinations are stored in paper format, and I collected all data from the oral exam fall 2012.

The data collected included time of day, grade, gender of both censors, which field of study the candidate belonged to, candidates' gender and age. In total, I collected data on 728

students from eight different programs. Observations where the candidate had a doctor's appointment, or did not show up was eliminated (n=22)[1].

The most important variable collected is the time of day. As previously explained there is a planned lunch for the censors at 11:50, each day, which last for one hour. Since I am looking for a lunch effect, this is of utmost importance, and makes my data comparable with the work of Danziger et al. (2011a) and Linder et al. (2014) which also used planned lunch as the main explanatory variable. I have no way of knowing whether the students taking the exam uses this break to have lunch themselves, I can only assume that they do their best to plan the day according to when they have their exam.

One clear difference between my dataset and both mentioned studies is that their observed outcome is a binary variable. In Danziger et al., the decision is "reject parole" or "approve parole" and in Linder et al., the decision is "prescribe antibiotics" and "do not prescribe antibiotics", whereas in my data set the decisions goes the full ordinal scale of A to F. This makes it somewhat difficult to compare results, as my data is ordinal. There are however, established numerical values for the ordinal scale, which is often used when dealing with grades.

4.1 Hypothesis and expectations

I expect the time of day to have an effect on the grade. The first candidate each day has his or her examination at 0800, followed by a new candidate 20 minutes after until 1200 when there is a 30 minute break before continuing at the same pace, until all candidates in the given student group have had their exam. Students are unequally distributed around lunch; most days there are more students before lunch than after lunch.

Before the thirty-minute food break, I expect the grades to be lower, than after the break, according to the theory of mental depletion.

[1] 11 had doctors certificates and 11 was no-shows

In the case of oral examinations, I believe that, as the day progresses towards lunch, the censors will be more inclined to give C's and D's. This is because on a scale from A to F those will be the grades we see the most, given that the grades are normally distributed.

After lunch, the censors will again be "able" to make the harder decisions, which I assume is giving the top grades, A and B, and the bad grades, E and F.

Hypothesis 1: Just before lunch, the grades will be lower, compared with just after.

Hypothesis 2: The censors will make the easiest choices before lunch, which is giving more C and D. This implies that the censors will make the harder choices after lunch, which is giving more A, B, E and F.

4.2 Analysis and results

Since the object of the current research is to examine whether a lunch effect is present in the exam results, I focused observations around lunch. I did this by only paying attention to the last four observation before lunch and the first four after lunch, "eliminating" all other observations. By doing it this way, I was able to compare the last four grades before to the first four after lunch.

After eliminating observations, I was left with 340 observations, focused around lunch.

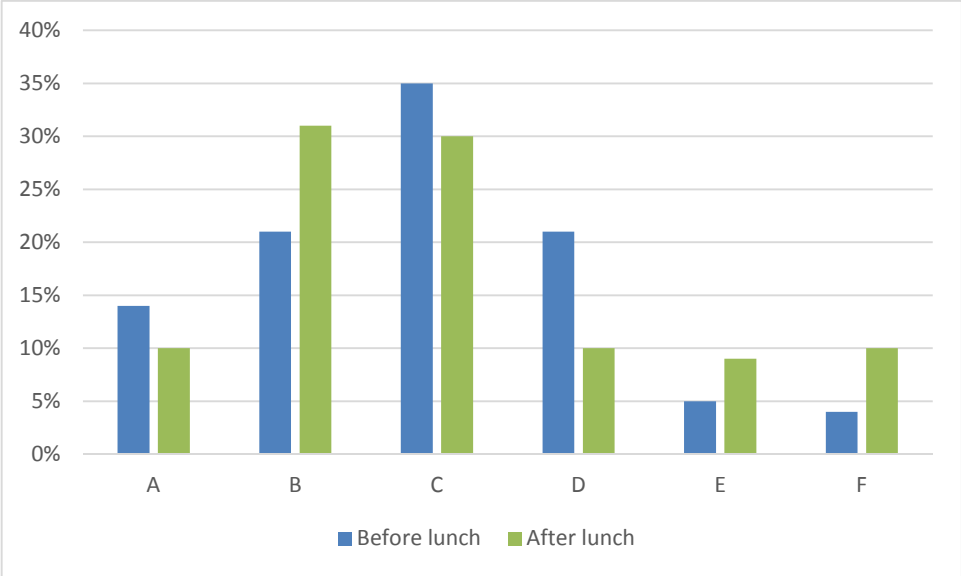


Figure 2 Distribution of grades around lunch for all students (n=340)

Based on the figure above alone, we can see that a lot more students are given a B after lunch than before and a lot more students are given a D before lunch than after. This seems in line with my first hypothesis that the censors will be giving lower grades just before lunch, as compared with after lunch.

4.2.1 Z-test of two population proportions

In order to see if these proportions are statistically significant from each other, I used a z-test for two population proportions. This test is used when you want to examine whether two populations or groups differ significantly from each other given a specific characteristic. The characteristic in this case is whether the candidate had the examination before, or after lunch.

I used the following formula to calculate the z-statistics[2]:

[2] http://www.socscistatistics.com/tests/ztest/Default.aspx

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}\,(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where, $\bar{p}_1$ and $\bar{p}_2$ is the proportion of the given grade in each treatment (before and after lunch), $\bar{p}$ is the proportion of the given grade combined. n_1 and n_2 denotes the total number of students in each treatment.

My null hypothesis is that the proportion of students getting a given grade before lunch is statistically the same as after lunch, and the alternative is that it is not. As it is somewhat unclear which way the lunch effect would affect the grade, I did a two-tail test, with significance level at 5%.

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 \neq 0$$

Presented below, is the results from the z-test when comparing grades before and after lunch.

Grade	Before lunch	After lunch	P-value
A	14 %	10 %	0.2983
B	21 %	31 %	0.0500*
C	35 %	30 %	0.3898
D	21 %	10 %	0.0071**
E	5 %	9 %	0.2077
F	4 %	10 %	0.03236*

Table 3: Proportions of grades before and after lunch, all students. * and ** denotes statistical significance at 5% and 1% respectively

As we can see the difference in proportions before and after lunch for the grades B (p = 0.05), D (p = 0.007) and F (p = 0.03) are all statistically significant (p<0.05). The results for B and F are in line with the expectations, as doing hard decisions require more mental capacity, so

they are avoided before lunch. There are a lot more students given a D, which is considered an "easy" decision before lunch, which also is in line with my expectations.

4.2.2 Robustness test

As presented above, there are more students given a B right after lunch, than right before lunch. There are also more students given a D right before than right after. These results seem in line with the theory of mental depletion. When the censors are doing continuous decisions throughout the day, they seem to make the "easy" decision right before lunch, which I propose is giving a C or a D. Right after lunch they will maybe pay more attention, and might be willing to give more consideration to their decision. In the case of the data presented, it looks like this result in better grades after lunch, given that more B's are given. The data also suggests that there are given more F's, which I would argue is due to giving an F is a "hard" decision, which requires more mental capacity. I assume that if a student presents little of the knowledge expected, it would be easier for the censor to give that student an E, rather than failing him, which the censor know has more implications for the student and the university, than letting him pass.

In order to examine the robustness of these results, I repeated the analysis, with the two largest groups of students. This was students from the faculty of Humaniora, samfunnsvitenskap og lærerutdannelse (HSL) and faculty of Psychology. If the results are robust, I expect the results to replicate for both these faculties. I chose these two groups, as they were by far the two largest groups. I did not repeat the analysis for the remaining groups, as the sample-size would be too small in order to obtain meaningful results.

HSL faculty

The students from the faculty of Humaniora, samfunnsvitenskap og lærerutdannelse (HSL), represent 214 students of the total 708, making them the largest group of students.

Once again, since the variable of interest is whether the censors have had their lunch, I concentrated the grades around lunch. The data presented in the figure below consists of the last four candidates before lunch, and the last four after lunch.

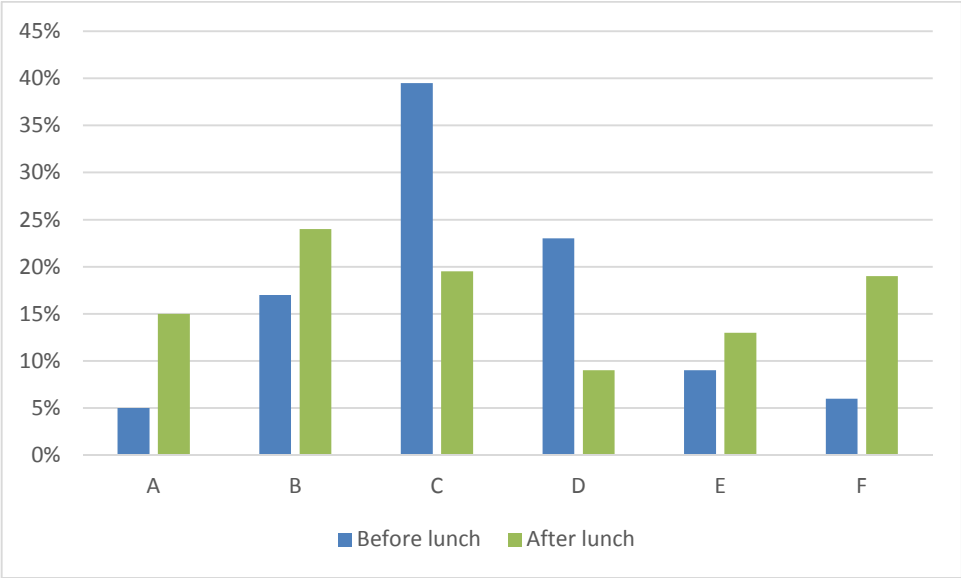


Figure 3 Distribution of grades around lunch for HSL students only (n=99) average of total grades (n=214).

By looking at the distribution, It seems in line with my hypothesis. More students are given A, B and F after lunch and more students are given C and D before lunch.

However merely looking at the data is not sufficient to say the proportion of grades are different from each other. In order to test this, I once again did the z-test for two population proportions, test statistics from a two-tail test as follows

Grade	Before lunch	After lunch	P-value
A	5 %	15 %	0.1164
B	17 %	24 %	0.3899
C	40 %	20 %	0.0300*
D	23 %	9 %	0.0601
E	9 %	13 %	0.5687
F	6 %	19 %	0.0349*

Table 4: Proportions before and after lunch, HSL students only. * and ** denotes statistical significance at 5% and 1% respectively

The results support the hypothesis that the censors do not want to take the hard choices before lunch, as more students are given the expected grade C before lunch than after. The effect on the HSL students seems to be big, as 39% of the students receive a C before lunch, whereas only 18% receive a C after lunch. There are also given more F's after lunch than before. Although not significant at 5% level, there seem to be more students given a D before lunch than after lunch, which is in line with the aggregated results. These results point in the direction that the censors are giving worse grades before lunch than after lunch.

Department of Psychology

The students from the psychology department of the university, makes up the second largest group of students taking the examen philosoficus exam (168 out of 708 students).

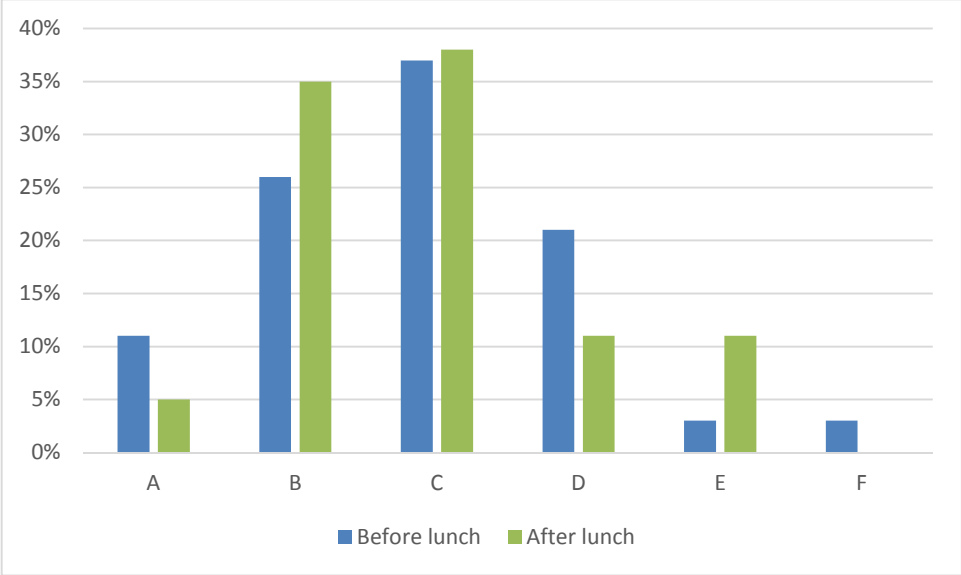


Figure 4 Distribution of grades around lunch, PSY students only (n=75) average of total grades (n=169).

By looking at the graph of distribution for the psychology students, the results does not seem as clear-cut. Here it looks like there are given more A, D and F before lunch, which is in contrast to my expectations. After lunch, it seems to fit more with my expectations, as there seem to be more B's and E's. However, the z-tests for two population proportions reveal that

nothing special is going on here, and none of the proportion of grades are significantly different from each other. While not statistically significant on the 5% level, there seem to be given more E's after lunch than after (10% level).

Grade	Before lunch	After lunch	P-value
A	11 %	5 %	0.2069
B	26 %	35 %	0.2038
C	37 %	38 %	0.4645
D	21 %	11 %	0.1132
E	3 %	11 %	0.0778
F	3 %	0 %	0.1603

Table 5: Proportions before and after lunch, PSY students only. * and ** denotes statistical significance at 5% and 1% respectively

By examining the data from the two groups on their own, it looks like the result is not as robust as expected. If it was the case that censors graded differently before and after lunch, I would expect the results to be somewhat the same in both groups. However, the two groups may not be entirely comparable, as students at the first year of psychology need close to an A average in order to be accepted at the vocational studies of psychology.

4.3 Pooled grades

When censors are faced with tough decisions, they will tend to make the easiest decision if they are mentally depleted. Presented graphically below is the results of pooling the grades C and D, which I consider the easy choice, and A, B, E and F, which I consider to be requiring more mental capacity. It looks like there is an overweight of "easy" grades C and D before lunch than after lunch and more of the harder grades after lunch.

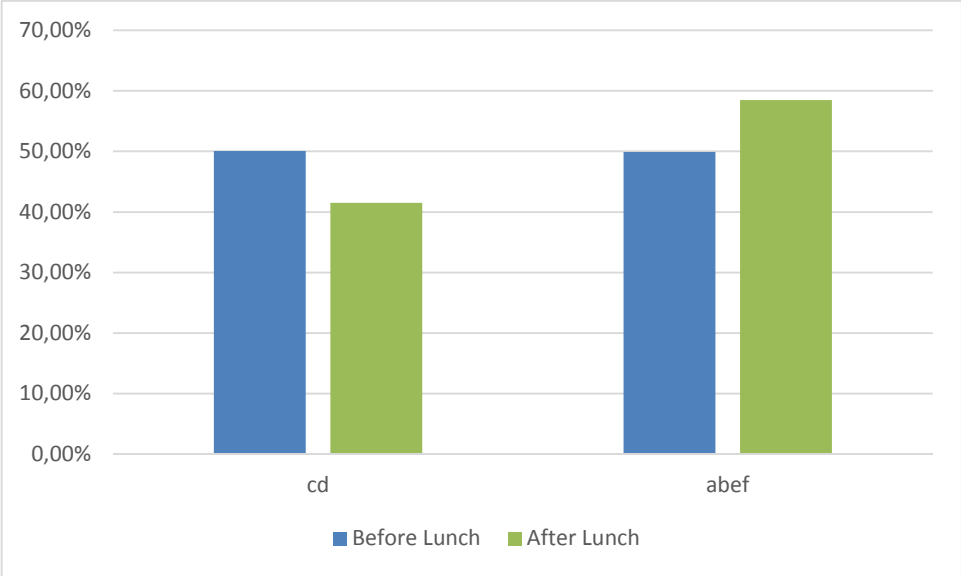


Figure 5: Proportion of pooled grades, before and after lunch for all students.

While it is impossible to conclude is these numbers are significantly different from each other just by looking at the graph, a z-test of two populations' proportions reveals that they are both significant at the 5% level.

Grade	Before lunch	After lunch	p-value
CD	50.10 %	41.50 %	0.034*
ABEF	49.90 %	58.50 %	0.034*

Table 6: Pooled grades before and after lunch, all students * denotes statistical significance at 5% level.

Given that giving a C or D is considered the "easy" decision, and A, B, E and F are "hard", these results gives a clear indication that a lunch-effect is present. There are a lot more students given the "hard" grades after lunch, compared with before, indicating that the hard decisions are done after lunch.

4.4 Ordinal scale method

Grades are ordinal scale data, which gives a rank order (A, B, C etc). The distance between A and B is not the same as the distance between a B and C. A is usually given to the top 10% of the class, while a C is given to about 40% of the class. However, the scale does not have a specific degree of difference between the grades. To test the relationship between grades before and after lunch, I tested the Spearman rank order coefficient, which is a non-parametric test designed to measure relationship between two ordinal scale variables (Corder and Foreman, 2009).

The average score of the last four candidates before lunch, and first four after lunch was collected. In cases where a candidate was missing (due to doctor's appointment etc), I took the average of three instead of four. My null hypothesis was that the spearman rho coefficient was zero and the alternative hypothesis that it was different from zero. I set the level of significance to 5%.

Testing this in Stata reveals that the spearman rho is not significant. (Rho=0.2982, p>0.05) Based on this statistics, I can state that there is no clear relationship between the grades before and after lunch.

Number of observations	=	41
Spearman's rho	=	0.2982
Test of H0: before and after are independent		
Prob > |t|	=	0.0583

Table 7: Output Spearman's rank coefficient test

While the spearman's rho coefficient is not statistically significant at the 5% level, it is significant on the 10% level, and indicates a moderate positive effect.

5. Discussion

Previous research on ego-depletion states that as decision-makers make repeated decisions throughout the day, they become mentally depleted, resulting in a tendency towards the status quo (Muraven and Baumeister, 2000a, Baumeister et al., 1998). Sticking to the status quo is regarded as the easy option, which do not require a lot of mental process. Research from behavioral economics describes the status quo effect as an implication of loss aversion (Kahneman et al., 1991), and these two may be closely related, as research have found that ego-depleted individuals tend to act more risk averse (De Langhe et al., 2008, Unger and Stahlberg, 2011).

Recent research implies that "expert" decision-makers, like judges (Danziger et al., 2011a) and clinicians (Linder et al., 2014) are also affected by ego-depletion, which results in making the "easy" decisions such as refusing parole or prescribing antibiotics.

The goal of the current research was to investigate whether the grading of oral exams was connected with the censors lunch break, and my hypothesis was that censors would tend to make the "easy" decision when ego-depleted. This hypothesis requires an assumption that giving a C or D is considered the "easy" decision, and giving A, B, E or F are "harder".

Data was collected from a quasi-field experiment, where students was assigned to either having their exam before or after lunch, in alphabetical order. This way of calling in the students, should be sufficient to say that the abilities of the students are distributed randomly before and after lunch. The subjects in my research are the censors, which grade the student shortly after the examination. This provided me the opportunity to determine if the grade was set before, or after the censors scheduled lunch-break. After collecting the data, I compared the four grades given just before lunch, to the four grades given just after lunch, by doing a z-test of two population proportions.

By comparing the grades given just before lunch to the grades given just after, I was able to focus the grades around lunch, as the goal was to examine "the lunch effect". The result, although not as robust, was in line with my expectation. Aggregated over all students, the results provided some evidence for a bias in the grading. There was statistically more students given a D ($p = 0.007$) before lunch, and more students given B ($p = 0.05$) and F ($p = 0.032$) after lunch. In order to test the robustness of my result, I rerun my calculation with smaller

groups, according to the students' field of research. I expected that if the result were robust, it would replicate in the two largest groups of students. This was however not the case, as the result for the largest student group (HSL) replicated in the sense that there was more C ($p = 0.03$) and D ($p = 0.06$) and more F after lunch ($p = 0.035$). In the case of Psychology students, these results did not replicate, but it revealed that there was given more E's after lunch, but only significant at the 10% level. The analysis was not repeated for the remaining groups of students, since the sample-size for these groups would have been too small.

A more interesting result occurred when pooling the grades, according to "easy" grades C and D, and "hard" grades A, B, E and F. The results revealed that there was significantly more C and D given before lunch ($p < 0.05$) and significantly more A, B, E and F after lunch ($p < 0.05$). The results from the pooled grades are in line with my expectations, and given that C and D is the "easy" decision, it is in line with both the presented previous studies on the lunch effect

6. Concluding remarks

The results from the current research is important, as it provides some support for the research done by Danziger et al. (2011a) and Linder et al. (2014). As previously stated, the current study is not an exact replication of either of these studies, but it examines the same effect. It seems probable that the censors are somewhat biased toward making the "easy" decision right before lunch. Restoring mental capacity during the lunch-break could make them able to do "harder" decisions after lunch. Again, I would like to stress, that as was the case with both the judges and clinicians, I had no way of recording the censor's state of mental capacity. The current research cannot state a *clear* causal relationship between the grades given and ego-depletion, in addition, there was no way to get data on whether or not the students was the cause, as they too could have been ego-depleted at the time of examination. However, the results suggests that something is happening around lunch.

More importantly, these results provides an addition to the literature on biased grading. It shows, although not entirely robust, that there may be a lunch-effect bias in university grading. Given that it exists, it should be addressed when examination guidelines are set. As

previously discussed, biased grading may result in a misallocation of resources. If grades are not based on performance alone, students could be lead to make inefficient choices as further education is concerned, which in turn could lead to a less efficient workforce (Kiss, 2013). If this is the case, it would have clear implications for the economy as a whole, as a suboptimal allocation of the workforce would result in lower levels of income than an optimal allocation.

Future research should seek to combine several methods in order to investigate a causal relationship between ego-depletion and grading. For instance, the censors' lunch break could be assigned to different times of the day. This could easily be argued to be less practical as examination is concerned, however, it would allow the researchers to randomize the censors into treatments, which is not the case in the current research. Other possibilities could be to measure the blood glucose levels of both student and censor at each examination, but this would also not be optimal, as quite possibly, this would induce extra stress on the student. However, similar environments could easily be created in the lab, although it would have less external validity.

7. References

ABELER, J. & NOSENZO, D. 2014. Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, 1-20.

ALMÅS, I., CAPPELEN, A. W., SØRENSEN, E. Ø. & TUNGODDEN, B. 2010. Fairness and the development of inequality acceptance. *Science,* 328, 1176-1178.

ARIELY, D., LOEWENSTEIN, G. & PRELEC, D. 2003. "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics,* 118, 73-106.

BAUMEISTER, R. F., BRATSLAVSKY, E., MURAVEN, M. & TICE, D. M. 1998. Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology,* 74, 1252.

BENABOU, R. & TIROLE, J. 2003. Intrinsic and extrinsic motivation. *The Review of Economic Studies,* 70, 489-520.

BÉNABOU, R. & TIROLE, J. 2000. Self-confidence and social interactions. National bureau of economic research.

BERGMAN, O., ELLINGSEN, T., JOHANNESSON, M. & SVENSSON, C. 2010. Anchoring and cognitive ability. *Economics Letters,* 107, 66-68.

BERNARD, M. E. 1979. Does sex role behavior influence the way teachers evaluate students? *Journal of Educational Psychology,* 71, 553.

CAMERER, C. F. & HOGARTH, R. M. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty,* 19, 7-42.

CAPPELEN, A. W., NYGAARD, K., SØRENSEN, E. Ø. & TUNGODDEN, B. 2011. Social preferences in the lab: A comparison of students and a representative population. CESifo working paper: Behavioural Economics.

CARD, D., DELLAVIGNA, S. & MALMENDIER, U. 2011. The role of theory in field experiments. National Bureau of Economic Research.

CHERRY, T. L., FRYKBLOM, P. & SHOGREN, J. F. 2002. Hardnose the dictator. *American Economic Review*, 1218-1221.

CLARKSON, J. J., HIRT, E. R., JIA, L. & ALEXANDER, M. B. 2010. When perception is more than reality: the effects of perceived versus actual resource depletion on self-regulatory behavior. *Journal of personality and social psychology,* 98, 29.

CORDER, G. W. & FOREMAN, D. I. 2009. *Nonparametric statistics for non-statisticians: a step-by-step approach*, John Wiley & Sons.

CROSON, R. 2005. The method of experimental economics. *International Negotiation,* 10, 131-148.

DANZIGER, S., LEVAV, J. & AVNAIM-PESSO, L. 2011a. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences,* 108, 6889-6892.

DANZIGER, S., LEVAV, J. & AVNAIM-PESSO, L. 2011b. Reply to Weinshall-Margel and Shapard: Extraneous factors in judicial decisions persist. *Proceedings of the National Academy of Sciences,* 108**,** E834-E834.
DE LANGHE, B., SWELDENS, S., OSSELAER, S. & TUK, M. A. 2008. The emotional information processing system is risk averse: Ego-depletion and investment behavior. ERIM Report Series Research in Management.
FLEMING, N. D. 1999. Biases in marking students' written work: quality. *Assessment matters in higher education: choosing and using diverse approaches***,** 83-92.
FRIEDMAN, D. 1994. *Experimental methods: A primer for economists*, Cambridge University Press.
GAILLIOT, M. T. & BAUMEISTER, R. F. 2007. The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review,* 11**,** 303-327.
GAILLIOT, M. T., BAUMEISTER, R. F., DEWALL, C. N., MANER, J. K., PLANT, E. A., TICE, D. M., BREWER, L. E. & SCHMEICHEL, B. J. 2007. Self-control relies on glucose as a limited energy source: willpower is more than a metaphor. *Journal of personality and social psychology,* 92**,** 325.
GIGERENZER, G. & GAISSMAIER, W. 2011. Heuristic decision making. *Annual review of psychology,* 62**,** 451-482.
HARRISON, G. W., LAU, M. I. & RUTSTRÖM, E. E. 2011. Theory, experimental design and econometrics are complementary (and so are lab and field experiments). *The Methods of Modern Experimental Economics*.
HARRISON, G. W. & LIST, J. A. 2004. Field experiments. *Journal of Economic Literature***,** 1009-1055.
HINNERICH, B. T., HÖGLIN, E. & JOHANNESSON, M. 2011. Are boys discriminated in Swedish high schools? *Economics of Education Review,* 30**,** 682-690.
HOFFMAN, E., MCCABE, K. & SMITH, V. L. 1996. Social distance and other-regarding behavior in dictator games. *The American Economic Review***,** 653-660.
INGRAM, J. K. 1888. *A history of political economy*, Macmillan.
JACOWITZ, K. E. & KAHNEMAN, D. 1995. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin,* 21**,** 1161-1166.
JOB, V., DWECK, C. S. & WALTON, G. M. 2010. Ego depletion—Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological science*.
JOB, V., WALTON, G. M., BERNECKER, K. & DWECK, C. S. 2013. Beliefs about willpower determine the impact of glucose on self-control. *Proceedings of the National Academy of Sciences,* 110**,** 14837-14842.
KAHNEMAN, D., KNETSCH, J. L. & THALER, R. H. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *The journal of economic perspectives,* 5**,** 193-206.

KAHNEMAN, D. & TVERSKY, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.
KENRICK, D. T. & GUTIERRES, S. E. 1980. Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology,* 38, 131.
KISS, D. 2013. Are immigrants and girls graded worse? Results of a matching approach. *Education Economics,* 21, 447-463.
KOUCHAKI, M. & SMITH, I. H. 2013. The Morning Morality Effect The Influence of Time of Day on Unethical Behavior. *Psychological science*, 0956797613498099.
KURZBAN, R. 2009. Does the brain consume additional glucose during self-control tasks? *Evolutionary psychology: an international journal of evolutionary approaches to psychology and behavior,* 8, 244-259.
LAKATOS, I. 1980. *Mathematics, Science and Epistemology: Volume 2, Philosophical Papers*, Cambridge University Press.
LANGE, F. & EGGERT, F. 2014. Sweet delusion. Glucose drinks fail to counteract ego depletion. *Appetite,* 75, 54-63.
LAVY, V. 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics,* 92, 2083-2105.
LEVITT, S. D. & LIST, J. A. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *The journal of economic perspectives*, 153-174.
LINDAHL, E. 2007a. Comparing teachers' assessments and national test results: evidence from sweden. Working Paper, IFAU-Institute for Labour Market Policy Evaluation.
LINDAHL, E. 2007b. Gender and ethnic interactions among teachers and students - evidence from Sweden. *University of Uppsala*.
LINDER, J. A., DOCTOR, J. N., FRIEDBERG, M. W., NIEVA, H. R., BIRKS, C., MEEKER, D. & FOX, C. 2014. Time of Day and the Decision to Prescribe Antibiotics.
LIST, J. A. 2011. Why economists should conduct field experiments and 14 tips for pulling one off. *The Journal of Economic Perspectives,* 25, 3-15.
LOEWENSTEIN, G. 1999. Experimental economics from the vantage‐point of behavioural economics. *The Economic Journal,* 109, 25-34.
MALOUFF, J. 2008. Bias in grading. *College Teaching,* 56, 191-192.
MECHTENBERG, L. 2009. Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The Review of Economic Studies,* 76, 1431-1459.
MILGRAM, S. 1963. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology,* 67, 371.
MURAVEN, M. & BAUMEISTER, R. F. 2000a. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological bulletin,* 126, 247.

MURAVEN, M. & BAUMEISTER, R. F. 2000b. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin,* 126**,** 247-259.
NISBETT, R. E. & WILSON, T. D. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology,* 35**,** 250.
PERSKY, J. 1995. Retrospectives: the ethology of homo economicus. *The Journal of Economic Perspectives***,** 221-231.
POPPER, K. R. 1954. *Conjectures and refutations*, Minumsa.
POPPER, K. R. 1959. The logic of scientific discovery.
SAMUELSON, P. & NORDHAUS, W. 1985. Principles of economics. New York: McGraw-Hill.
SAMUELSON, P. A. 1937. A note on measurement of utility. *The Review of Economic Studies,* 4**,** 155-161.
SAMUELSON, W. & ZECKHAUSER, R. 1988. Status quo bias in decision making. *Journal of risk and uncertainty,* 1**,** 7-59.
SAVAGE, L. 1972. *The foundations of statistics*, DoverPublications. com.
STROOP, J. R. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology,* 18**,** 643.
THALER, R. H., TVERSKY, A., KAHNEMAN, D. & SCHWARTZ, A. 1997. The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics,* 112**,** 647-661.
TICE, D. M., BAUMEISTER, R. F., SHMUELI, D. & MURAVEN, M. 2007. Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology,* 43**,** 379-384.
TVERSKY, A. & KAHNEMAN, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology,* 5**,** 207-232.
TVERSKY, A. & KAHNEMAN, D. 1974. Judgment under uncertainty: Heuristics and biases. *science,* 185**,** 1124-1131.
TVERSKY, A. & KAHNEMAN, D. 1991. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics,* 106**,** 1039-1061.
TYLER, J. M. & BURNS, K. C. 2008. After depletion: The replenishment of the self's regulatory resources. *Self and Identity,* 7**,** 305-321.
UNGER, A. & STAHLBERG, D. 2011. Ego-depletion and risk behavior: Too exhausted to take a risk. *Social Psychology,* 42**,** 28.
UUSKARTANO, K. J. B., SVETLANA S 2013. Challenges in Experimental Economics: Discussions based on a field study among fishermen. *Master Thesis, University of Tromsø*.
VARIAN, H. R. & NORTON, W. 1992. *Microeconomic analysis*, Norton New York.
WEICHSELBAUMER, D. & WINTER‐EBMER, R. 2005. A Meta‐Analysis of the International Gender Wage Gap. *Journal of Economic Surveys,* 19**,** 479-511.

WEINSHALL-MARGEL, K. & SHAPARD, J. 2011. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences,* 108**,** E833-E833.
ZIZZO, D. J. 2010. Experimenter demand effects in economic experiments. *Experimental Economics,* 13**,** 75-98.