LAB EXPERIMENT FOR A NEW COMPENSATION MODEL

A REVIEW OF THE LITERATURE

DRAFT NOTE

AUTHORS
NILS VERHEUVEL, THOMAS ONGENAE, ASTRID LENSINK, FLORIAN VAN LEEUWEN

COMMISSIONED BY
REWARD VALUE

AMSTERDAM, NOVEMBER 21

SEO note <reference number>

Roeterstraat 29
1018 WB, Amsterdam
The Netherlands

+31 20 525 1630
secretariaat@seo.nl
www.seo.nl/en/

1 Introduction

Companies aim to create value to its stakeholders. Increasingly, these stakeholders are not limited to shareholders, but also include society at large and the environment. The task of the chief executive officer (CEO) is to steer the company towards value creation. A compensation package is meant to incentivize the CEO to create value for stakeholders. However, current practice shows that most CEOs are rewarded only for financial value creation, and often only focused on the short run. Reward Value Foundation's mission is to change executive compensation models so that companies can be a catalyst for change and a shift towards sustainable long-term value creation.

The new compensation model should have a scientific basis. To this end, Reward Value has commissioned SEO Amsterdam Economics to conduct research on elements for a new compensation model. The outline of the new model is described in the green paper "Rewarding stakeholder long-term value creation."[1] SEO has conducted an extensive literature review and data analysis to construct elements of a new compensation model.

This compensation model is tested in an online experiment, which was conducted in early 2021. The experiment showed that participants, who take the role of a CEO of a fictional company, respond to incentives. In the experiment, participants can decide between short-term production and investment in R&D for long-term production. The status quo is that the bonus for short-term production is higher than the bonus for long-term production. The experiment also includes a compensation model that gives an equally high reward to long-term production, or that introduces risk in the pay-off of the experiment. Another experiment involves the choice between cheap, but environmental-unfriendly production, or more expensive, but environmental-friendly production. The results of the experiment shows that changes in the bonus structure can affect investment choices. An important finding was that the results differ based on the personality traits of the participants.

This goal of the current literature review is to improve the experimental design to assess the effects of various executive compensation models on firms' sustainable long-term value creation. A new research agenda builds upon the existing work and expands the experiment to include more behavioural insights on CEO personality and preferences, as well as a better understanding of what drives the differences in outcomes and how to ensure that this effect is a causal estimate. The first step is to distil lessons from the academic literature.

This note describes the findings from the literature review on previous experiments, CEO behaviour and preferences, and methodology of experiments. SEO has studied around ninety-six academic articles, divided into three main topics. First, we describe previous experiments on executive compensation models from the literature and discuss the lessons for our experiment. The next chapter describes which CEO personality traits affect firm outcomes and what is the underlying mechanism. Additionally, it points towards improvements in our current experimental design. The final chapter describes the methodology behind experimental designs and describes several lessons to ensure the experiment allows for an identification of causal effects. We conclude by listing paths to develop the current experimental design.

[1] https://www.rewardvalue.org/research/completed-research/

2 Lessons from previous experiments

Executive compensation models and their effects on executive behaviour and firm performance have been a topic of interest in the literature. The relation between several types of payment structures and performance has been tested in many laboratory experiments (Bruner et al., 2008; Ariely et al., 2009; Agranov & Tergiman, 2013; Harris et al., 2018). Bruner et al., (2008) used a survey setup to investigate the potential downsides of equity-based compensation, and they found an increase in fraudulent behaviour. The experimental setup was similar for Agranov & Tergiman (2013), who investigated different payment schemes, and Harris et al. (2018), who examine how bonus caps and malus affect individuals' choices of risk and effort. Both employed a survey-based experiment.

There also have been numerous field experiments showing that performance-based compensation could lead to an increase in productivity, but there is no consensus in the literature on this (Camerer, Hogart, 1999; Lazear, 2000; Shearer, 2004; Bandiera et al., 2007; Gielen et al., 2010; Ottersen, 2016; Iqbal et al., 2019).

2.1 General lessons

The three main lessons from previous experiments concern sample section, shareholder views, and performance-based compensation. They are explained in successive order.

2.1.1 Sample Selection

A great part of experiments in social sciences is performed with the participant group consisting of undergraduate students. This group is often chosen due to its accessibility and low costs. Undergraduates are, however, often not the sample group of interest. In many cases, an ideal sample would be less homogeneous. A solution is to select a sample group similar to the group of interest, e.g. business students when the researcher is interested in executives.

Andersen and Lau (2018) offer an alternative, they use Amazon Mechanical Turkers. These are crowd workers who perform discrete on-demand tasks for a set rate. The "Turkers" have become a cheap option to quickly recruit many respondents (Berinsky et al., 2012; Paolacci et al., 2010). The benefit of using Turkers is that they provide a more diverse sample size, compared to most undergraduate samples (Huff and Tingley, 2015). It can, however, be difficult to select a representative group of future CEOs as these are likely to not be crowded workers.

Another option is Prolific, the platform offers a diverse database and is specifically made for research purposes. Subjects can be screened in a range of dimensions, making it possible to acquire a sample group with specific characteristics, for example higher education or management experience. The platform combines good recruitment standards at a reasonable cost (Palan & Schitter, 2018). In a range of experiments, participants from Prolific gave more reliable answers than Turkers and university students. This was measured by attention check-questions and in how well results from established experiment could be replicated (Peer et al., 2017).

2.1.2 Shareholder views

CEOs of publicly traded companies know that the results of the company will influence shareholders' voting behaviour and in turn the executive's compensation. CEOs thus must keep in mind that they might only be able to keep their position by not taking too many risks. Shareholders seem to only respond negatively to high CEO rewards when firm performance is low, due to loss aversion (Krause et al., 2014). However, for some investments in environmental and social causes risks are more tolerated by shareholders.

CEOs can expect the support of shareholders in making sustainable investments before and after they are implemented, even though the return on investment might be low. The pressure for companies and pension funds to take social and environmental views into consideration is rising. These institutions' function is to generate value for the stakeholders, which includes the shareholders. It is thus important to know if shareholders agree with the push to think more about other stakeholders. In two surveys run by Bauer et al. (2021), they answered this question by asking participants' views on investment decisions whose pension fund gave that grants its members a real vote on the sustainable-investment policy. The first survey showed that 68 percent supported more sustainable investment, even though the expected value was negative compared to non-sustainable investments. The second experiment determined that support remained strong after the pension fund had implemented the new sustainable-investment policy.

Social values of companies are becoming a subject where shareholders are staring to express their concerns increasingly in the form of shareholder proposals. Firms targeted by social shareholder proposals have higher scores on employee wellbeing and human rights performance (Eding, Scholtens, 2017). In some cases, picking a manager who spent resources on social causes, rather than strive for short-term maximization for shareholders profit, can be more profitable in the long run due to the increase in employees' performance (Kajackaite, Sliwka, 2020). A famous example of this practice is the 105 percent pay raise Henry Ford gave to 90 percent of his employees working at his automobile plant in 1914. The sustained productivity increase that followed was between 40 and 70 percent (Daniel et al., 1987).

2.1.3 Performance-based compensation

Companies can choose from wide a variety of payment structures; these usually consist of one or more parts. The parts of the payment structure can generally be divided into the following three categories: non-variable cash compensation (fixed salary), performance-based cash compensation, and performance-based equity or option compensation.

The general problem with performance-based compensation is that the indicators used must be easy to measure and this thus far has prohibited the widespread use of social and environmental indicators. The next two subsections explain different forms of performance compensation and the third subsection focuses specifically on the compensation of CEOs.

Performance-based cash compensation

Economic experiments show that most types of payments structures have both up- and downsides compared to non-variable cash compensation, where the upside consists of increased effort and the downsides consist of fraud and excessive risk-taking. Ariely et al. (2009) concluded in several independent experiments that performance-based variable cash compensation can have a positive effect on effort but that raising incentives beyond a certain level may produce supra-optimal levels of arousal and hence decrease performance, which is in line with the Yerkes–Dodson law.

In a field experiment with auto windshield installers, the payment schema was switched from an hourly rate to a performance-based rate, where the workers were paid a set amount for each windshield that they installed. The result saws that productivity almost immediately increased by 44 percent (Lazear, 2000). There have been other experiments showing the same results, not only for blue-collar workers but also for managers (Shearer, 2004; Bandiera et al., 2007; Gielen et al., 2010; Iqbal, 2016).

In another field experiment by Bun and Huberts (2018), it was shown that lowering performance-based cash compensation paired with an increase in fixed salary decreased productivity on average. Their study analysed a change in payment structure of a large marketing firm. The changes in pay structure caused, on average, a reduction of more than 7 percent in sales.

Performance-based cash compensation can also have a positive effect on the risk assessment of employees if used correctly. Volume incentives can distort the assessment of credit risk, for non-experienced as well as experienced personnel at financial institutions (Cole et al., 2015). Incentives that reward performance and penalize failure had a positive effect on risk assessment.

There is no clear census in the literature on whether performance-based cash compensation increases productivity, Camerer and Hogart (1999) argue in a meta-study of seventy-four different experiments that performance-based cash financial incentives have no effect on mean performance, but rather decrease the variance in the performance of the participants.

Performance-based equity compensation

Performance-based equity compensation is a common feature but next to incentivizing effort, it can also lead to more fraud because the sensitivity of CEO wealth increases with performance (Bruner et al., 2008). Equity is often seen as a measure to link the behaviour of executives to the long-term value creation of the company. It is common for companies to base equity pay on the company's relative performance over a three-to-five-year period. The awarded equity is usually locked up for a minimum of five years. Payment schemes that reward employees for realized investment returns but do not penalize negative returns, encourage excessive risk-taking. Two commonly used measures to combat the potential forms of fraud and exercise risk-taking are: Bonus caps and claw-back measures. Bonus caps create a ceiling for the variable compensation, and claw-back measures make it possible to reclaim bonusses that were awarded in the past. A disadvantage to the bonus cap is it can potentially cause a drop in effort after the maximum bonus is reached. This is less the case for the claw-back measures (Harris et al, 2018).

Performance-based compensation of CEOs

There have been theoretical models on improving the performance-based compensation of CEOs, some of which have been empirically assessed. An example is to base the variable compensation on not only stock price but also the price of debt, as proposed by Bolton et al. (2011). This should combat the excessive risk-taking, as the price of debt rises with an increase in risk and thus decreases compensation. They provided an empirical analysis that including debt into the compensation of executives, and they concluded that their compensation model is believed by the market to reduce the risk for financial institutions. Edmans & Liu (2011) proposed a similar theoretical model that granted managers not only equity but also debt. Another option to improve the current pay structure is to set the performance measure relative to the performance of other companies (compensation is negatively correlated with output of other companies). This can incentivize the effort of executives (Agranov & Tergiman, 2013). The findings from these studies can be used to improve the compensation model in our experiment.

2.2 Incorporating lessons in our experiment

For the new experiment, the lessons from the sample section can be used to create a representative sample for our CEO population. The insight the literature provided about shareholder views is necessary for providing background information to our participants. The lessons learned from different compensation models can be used to optimize our own compensation model.

2.2.1 Sample Selection

It seems that the easiest way to obtain a representative sample is to use the platform Prolific. Their platform allows for selecting participants in management functions. If this is not possible, then MSc students from management and finance programs can be used as the sample representative of future directors (Harris et al., 2018).

2.2.2 Shareholder view

Participants in the experiment should know the consequences of their actions. Our experiment could provide participants with the information that their investments in environmental and social causes are appreciated by shareholders, even though it might lead to a lower return on their investment. This could be a point of analysis; how do participants allocate resources to production when they do receive this information and how do they allocate resources when they do not receive this information. Participants could also be provided the information that shareholders do not like excessive risk-taking. Another option is that participants could lose their "job" between rounds if losses are large, this would reflect a more realistic scenario.

2.2.3 Performance-based compensation

In our first experiment it was determined that certain performance-based compensation measures could nudge participants to socially desired outcomes. In the next experiment the goal is to improve the nudges to create an even better supported compensation model.

The lessons from the literature and experiment on performance-based compensation can be incorporated by including the findings in the setup of our experimental setup. Claw-back measures can be included in the experimental setup, as was done by Harris et al. (2018). The experiment had participants make investments and they were paid a bonus proportional to the asset return only if the return exceeded a predefined threshold in the first period and conditional on if the project succeeds in the second period. The probability of making the targets was the same for both periods and stated with the list of investments.

Another option is to provide the participants with an option to invest not only in environmental investments but also in social investments. The return in the first period would be zero, and the production would increase in later periods due to an increase in productivity of "employees." This works better if there are more than two periods.
The risk treatment could be improved in two ways, by including a riskier investment option and making the compensation dependent on the risk participants take. In the previous experiment participants in the risk treatment had the option to choose a risky investment which had two outcomes with a positive return on investment. To make the investment decision more realistic the next experiment could include a risky investment consisting of one positive outcome and one negative outcome. Another option is to make the compensation dependent on the risks a participant takes. We could test if participants decrease the amount of risk they take if the compensation would be lowered for excessive risk-taking.

3 CEO behaviour and preferences

Personality traits have an influence on how CEOs respond to a remuneration policy. As personality traits help explain consistency in behaviour and explain important life outcomes (Roberts et al., 2007), it is important to consider these when assessing the remuneration policy. Different remuneration schemes will have different effects, and the size of this effect is not homogenous for all CEOs. This chapter describes which CEO personality traits affect firm outcomes and describes the underlying mechanisms. The first section describes the definition of a CEO. The second section lists the main findings and methods of relevant literature on CEO personality traits. Hereby, a distinction is made between 'demographic characteristics' and 'soft skills'. The third section explains how these traits can be incorporated in the experiment.

3.1 What is a CEO?

Publicly listed companies generally have many and frequently changing shareholders. These owners are not able to manage the company simultaneously. In addition, they often lack the skills to effectively manage the company. Therefore, they delegate the daily management of the company to (a board of) executives. The chief executive officer (CEO) is the highest-ranking executive in a company and serves a unique role in the organization as he/she determines the firm's direction, the relationship with key stakeholders and the general reputation. Because of their prominent influence on the organization, the CEOs personality characteristics are not solely visible in their personal preferences and behaviours, but also reflected in the strategies, structures and performances of their organization (Resick et al., 2009).

A principal-agent problem arises between the shareholders and the executive(s) - Mirlees (1999) and Holmstrom (2017). Shareholders must find the right executive(s) (which is difficult because of adverse selection), prevent opportunistic behaviour by the executives (who might suffer from moral hazard) and verify the performance (which is costly). In practice, there does not seem to be a perfect solution for the principal-agent problem. An option is the monitoring of the executive. In addition to monitoring, (financial) incentives are created in such a way that they align in the best way possible with the interest of the shareholders (Ter Weel, Witteman & Verheuvel, 2019).

3.2 Which CEO personality traits affect firm outcomes?

There are several types of personality traits affecting firm outcomes. Firstly, there are demographic characteristics that can easily be measured like age, education and gender. Secondly, there are soft skills. Examples are the Big Five characteristics: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. The characteristics of CEOs and the differences among them have been widely studied in both the academic and the business literature. In general, the literature states that personality traits of CEOs have a significant effect on activities within the organization and its performance, and thus proves that individual top executives do matter (Garcés-Galdeano & García-Olaverri, 2019). In this section, the demographic characteristics are discussed firstly. Secondly, the soft skills are discussed.

3.2.1 Demographic personality traits

The literature shows that the demographic characteristics of CEOs influence firm outcomes. To begin, the effect of gender on CEO pay and performance have been researched increasingly in the past literature. Tenure, age, experience, and education are discussed accordingly.

Gender

The empirical evidence regarding gender and CEO compensation is mixed. While some papers find that the compensation of female and male CEOs is not significantly different, others do document that female CEOs receive significant lower compensation. On the one hand Edmans et al. (2017) conclude that there is no significant difference in the yearly salary of males and females, after controlling for firm and individual characteristics. However, they do report that female CEOs only make up a very small fraction (2.5%) of their sample and moreover that female CEOs tend to run smaller firms. Also, Jordan et al. (2007), Ham et al. (2018), and Bugeja et al. (2012) do not report a significant different between female and male CEO pay. On the other hand, Elkinawy & Stater (2011) find that, after controlling for several characteristics, the salaries of female executives are around 5 percent lower than those for male executives. And thus, they do document an effect of gender on CEO pay.

Additionally, gender is often incorporated as a control variable in models estimating firm performance. In most cases, no significant effect of gender is found (Ham et al. (2018); Cole et al. (2015); Gilley et al. (2008)). Huang (2013) looks from a different perspective and concludes that a firms' Corporate Social Responsibility (CSR) performance, which is measured as the consistency of their CSR ranking, is associated with the CEOs' characteristics. While other papers find insignificant effects of gender, Huang (2013) concludes that male CEOs influence a firm's CSR performance positively. In his article he investigates the link between CEO demographic characteristics and consistency in CSR performance among firms. Although this is not directly related to firm performance, it does provide information regarding the new remuneration policy that is closely related to the (Environmental, Social & Governance) ESG criteria.

Tenure

Next to gender, Huang (2013) concludes that also the CEO tenure has a positive effect on firm performance. Anitia et al. (2010) describe the decline of the average CEO tenure from eight to four years in the last 20 years. Consequently, the pressure on CEOs to deliver quick results has spiked dramatically. In line with this, they conclude that a shorter CEO horizon is associated with more agency costs, higher levels of information risk and consequently lower firm valuation. This is also in accordance with the findings of Barker & Mueller (2002) who conclude that relative R&D spending increases with a longer CEO tenure. On the other hand, Ham et al. (2018), show that tenure does not influence firm investment and firm performance significantly. However, concluding from above, it is common practice in the literature to include tenure when analysing firm performance.

Age, education, and experience

Furthermore, age, education and experience are found to affect firm outcomes or are included in the model as a control variable. Next to gender and tenure, Ham et al. (2018) and Huang (2013) included age as a control variable, but it was found insignificant in both articles. Kaplan et al. (2012) studies a wide range of CEO candidates on the relationship between CEO characteristics and performance. R&D spending is higher when the CEO is younger, when they have more career experience in marketing and/or engineering/R&D and have education related to advanced science (Barket & Mueller, 2002). Garcés-Galdeano & García-Olaverri (2019) find that young, well-educated CEOs with external experience will improve innovative performance and the growth of the company. Cole et al. (2015) control for age, experience, rank, and the education in their model. Finally, Gilley et al. (2008) include survey questions in their written questionaries sought data such as age, the title, level within the organization, type of industry and the gender.

The literature discussed above, shows us that the most important characteristics that should be included in the model are gender, age (in years), tenure (in years), education, and experience. Furthermore, we could consider including rank, title, and the type of industry. Although a significant effect of the above-mentioned characteristics is not always found, it is can still be important to include them as control variables.

3.2.2 Soft skills

Soft skills, like the Big Five characteristics, overconfidence, narcissism, altruism, and risk aversion, also have an important effect on firm outcomes. However, these skills are harder to quantify and to incorporate in the model than the demographic characteristics. CEO's soft skills can affect firm outcomes and these traits can for example be investigated with the Big Five personality traits: agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. This section first discusses the Big Five personality traits. Since the Big Five are not the only personality traits that can have an influence on CEO behaviour, we discuss overconfidence, narcissism, and risk aversion in the second section.

The Big Five

Gow et al. (2016) show that the Big Five personality measures have a strong out-of-sample predictive performance (applying this method on different samples also can predict firm performance) and are stable personality traits for individuals over time. Furthermore, the Big Five personality traits are associated with choices for finance and investment and thereby for firm operating performance. The Big Five framework, created by Goldberg (1993), has attained a central place in phycological research. Personality psychology views these five personality traits as the "pattern of thoughts, feelings and behaviours that reflect the tendency to respond in certain ways in certain circumstances" (Roberts, 2009, p.140). Table 3.1 elaborates on the five personality traits.

Table 3.1 Big Five personality traits

Personality trait	Description
Agreeableness	Agreeableness refers to how people treat relationships with other people. It focuses on the people's orientation and interaction with others (Ackerman, 2017). The people that score high on agreeableness can be described as soft-hearted, trusting, and well-liked. Moreover, they are sensitive to the needs of others and are helpful and cooperative. Other people regard them as trustworthy and altruistic. The people that score low may behave suspicious, manipulative, and uncooperative. Others can regard them as less trustful.
Conscientiousness	Conscientiousness describes a person's ability to regulate their impulse control to engage in goal-directed behaviours (Grohol, 2019). A high score is correlated with carefulness, organization, disciplined, focus on detail and thoughtful. Tasks are likely to be completed and goals to be achieved. A low score indicates disorganizations, impulsiveness, and careless behaviour. This generates more difficulties in completing tasks and achieving goals.
Extraversion	Extraversion indicates how outgoing and social oriented a person is. When a person scores high, they enjoy being with others and like to participate in social gatherings. They thrive in social situations and feel comfortable to state their opinion. On the other side of the spectrum are the introverts, who are in general more reserved and quieter. Social interactions are more regarded as fatiguing and working alone is often a comforting setting (Lim, 2020).

Personality trait	Description
Neuroticism	Neuroticism refers to the ability to remain stable and balanced and thus to emotional stability. It also includes the extent to which a person can deal with negative emotions. If a person scores high on neuroticism, they tend to feel anxious, insecure, and experience larger shifts in mood. People that score low on neuroticism are described as calm, confident and, emotionally stable. In general, they have a higher self-esteem and are more resilient (Lim, 2020).
Openness	Openness relates to the willingness of people to engage in new activities. It is also related to the ability to think outside the box. On the high end of openness are people that can be described as curious, creative, and open to try new things. On the low end are people that dislike change, prefer routine work and like predictability (Lim, 2020).

O'Reilly et al. (2014) link the personality of CEOs to organizational culture. They find that organizational culture affects a broad set of organizational outcomes, such as financial performance, reputation, analysts' stock recommendations, and the attitudes of employees. They assess the personality of CEOs with the Big Five factor model. They obtain a score for each CEO by asking the employees of the firm to rate their CEO. This score was then included in the model. O'Reilly et al. (2014) are not alone when it comes to incorporating the Big Five characteristics in their model. Gow et al. (2016) state that the Big Five characteristics are associated with financial choices, investment choices and firm operating performance, linking their research to O'Reilly et al. (2014). For example, openness is positively related to R&D investments, whereas conscientiousness is negatively related to growth. Rather than attempting to measure personality traits using questionnaires or interviews, they create a method using the linguistic features exhibited by CEOs during conference calls. This method is highly appropriate to obtain the characteristics for a large sample. For smaller samples interviews and questionnaires of executives are feasible. Additionally, Bono & Judge (2004) consider a meta-analysis of the Big Five personality traits. Despite generally weak result of other traits, they conclude that extraversion could be an important trait when predicting and understanding transformational and transactional leadership. Hence, it seems that extraversion indicates robust relations with leadership outcomes.

Overconfidence, narcissism, and risk aversion

Several other soft skills, like overconfidence, narcissism, risk aversion and altruism, are also linked to firm outcomes. These personality traits are discussed below.

Overconfidence

Overconfidence in CEO behaviour can be defined as the overestimation of the value a manager/CEO believes he or she can create. It can influence the degree of risk perception and risk taking. Malmendier and Tate (2005) propose the systematic tendency to hold options longer before exercise as a measure of overconfidence. Overconfident executives namely overestimate the future performance of their firms and hence are more eager to hold options since they expect to profit from higher stock prices. This has become the most common approach to estimate overconfidence, it was also further applied by Malmendier and Tate (2015). Other options to measure over confidence include earnings forecasts, survey responses, or psychometric tests.

Narcissism

Narcissism, in the context of a personality trait, can affect firm investment and performance. It is positively linked to overinvestment and lower productivity (Ham et al., 2018). Since there seems to be a strong positive correlation between signature size and narcissism, Ham et al. (2018) propose a method of CEO signature size to examine the relation between the CEO's narcissism and firm investment. It is an informative method since it does not require study participants to answer direct questionaries about their personality, and because participants are likely to be unaware that their ego would affect something as simple as their signature. Narcissism has been identified as a stable personality trait that can be measured with personality assessment tools.

Risk aversion

Risk aversion is the tendency to prefer results with little uncertainty and can influence firm decisions, as investments often include a degree of risk. People that are more risk-averse tend to prefer more conservative investments to higher-yielding but less certain ones. Risk preferences can be measured by using a question framework (based on a gamble choice and a guaranteed payoff) from Bruner et al. (2008) for which they provided a seven-part classification system from very risk loving to very risk adverse. This lottery approach is widely used to measure risk aversion. Another example is Harrison & Rutström (2008), who measure risk aversion in a laboratory lottery setting.

Altruism

Altruism shows a desire to help others and is a sign of a lack of selfishness. Haynes et al. (2015) conclude that managerial altruism in general leads to a focus on longer-term decisions and long-term firm performance. However, they find that extreme altruism will worsen firm performance and hence a balance between altruism and self-interest leads to greater firm outcomes. These findings are in line with the previous experiment, where we found that altruistic respondents preferred environmental-friendly production even without receiving a higher bonus for it. The higher bonusses mainly influenced low altruism respondents, as they increased environmental-friendly production.

3.3 Incorporating personality characteristics in our experiment

For our experiment, it is important to include the demographic characteristics and soft skills in our model. The first part describes the incorporation of demographic characteristics and the second part the incorporation of soft skills.

3.3.1 Demographic characteristics

We can improve our experimental design by incorporating demographic characteristics by adding straightforward questions in the experiment. Although papers show mixed results regarding the characteristics affecting firm outcomes, it is suggested to include the mentioned characteristics: age, tenure, experience, education, and gender in de model. This data can easily be incorporated into the model as a vector of observed characteristics and function as control variables. When included in the model, it can of course be tested whether these have a significant effect or not.

3.3.2 Soft skills

As discussed, several measures are proposed to measure soft skills. To estimate CEO overconfidence the most common empirical approach was first established by Malmendier and Tate (2005). Building on their logic, Malmendier and Tate (2015) proposed the systematic tendency to hold options longer before exercise as a measure of overconfidence. The same holds for narcissism, which can be measured using signature sign, as proposed by Ham et al. (2018). Since these methods are complicated and time consuming for our research, we do not control for we overconfidence and narcissism. The Big Five characteristics, risk aversion and altruism are less complex to measure and can be incorporated in our model. The approach is described below.

To measure the Big Five characteristics, O'Reilly et al. (2014) asked employees of the firm to rate their CEO. The outcomes of these surveys were used to assign a CEO to certain characteristics, which were included in the model. The paper of Gow et al. (2016) build upon the article of O'Reilly (2014), they develop their measures using the linguistic features exhibited by CEOs during conference calls. For our experiment it is the most convenient to measure the Big Five personality traits based on interviews and questionnaires. For the lab-experiment we include a limited set of questions from the Big-five personality test to determine these personality traits. For the field-experiment, we can ask employees as well as the CEO himself/herself to fill in a survey with a set of questions of the Big-five personality test.

Risk preferences, and hence risk aversion, can be measured by using a question framework with a seven-part classification system from very risk loving to very risk adverse. Participants were giving a table (like Table 3.2) where they can choice between option A and option B. Option A is always a gamble, where option A is a gamble with options ranging from a very risky to a very safe gamble. Option B always a save fixed pay-out. A dummy can be included if a participant exhibits risk-averse preferences (Bruner et al., 2008).

Table 3.2 Table to measure risk aversion

Option	Gamble	Guaranteed payoff
	Choice A	Choice B
1	10% chance of 1000$ and 90% chance of 0$	500 $ guaranteed
2	20% chance of 1000$ and 80% chance of 0$	500 $ guaranteed
3	30% chance of 1000$ and 70% chance of 0$	500 $ guaranteed
4	40% chance of 1000$ and 60% chance of 0$	500 $ guaranteed
5	50% chance of 1000$ and 50% chance of 0$	500 $ guaranteed
6	60% chance of 1000$ and 40% chance of 0$	500 $ guaranteed
7	70% chance of 1000$ and 30% chance of 0$	500 $ guaranteed
8	80% chance of 1000$ and 20% chance of 0$	500 $ guaranteed
9	90% chance of 1000$ and 10% chance of 0$	500 $ guaranteed
10	100% chance of 1000$ and 0% chance of 0$	500 $ guaranteed

Source: Bruner et al. (2008)

Lastly, altruism was included in our previous experiment and could be included again. Altruism can be measured by asking questions in the survey regarding the distribution of obtained money from a lottery. Cole et al. (2015) ask the question what an individual would do when winning a lottery of € 100,000, whether the individual would keep it all for himself/herself of divide the money with family and friends. For robustness, we could incorporate different amounts of the lottery (like € 1,000, € 10,000, € 100,000, € 1000.000).

4 Methodology

It is of great importance to have the right methodology in place before setting up an economic experiment. It can namely contribute to the objectivity of the research. In the next sections, methodological aspects such as sample size calculations, ways of estimating treatment effects and randomization are discussed. The chapter starts by explaining the main types of experiments. Then, section 4.2introduces the concept of sample size calculations and further explained. Section 4.3 discusses the issue of what to look for when calculating statistical power. The following section takes a closer look at randomization and section 4.5 discusses several methods for estimating treatment effects. Finally, in section 4.6 the methodological lessons for our experiment are further examined.

4.1 Type of experiments

An experiment is a setting in which a certain group receives a so-called 'treatment', a factor in which the researcher is interested, and another group does not receive 'treatment', this is the control group. Comparing the outcomes of the treatment and control group leads to a causal estimate of the treatment variable of interest. There are three types of economic experiments, namely: laboratory, field and natural. In this chapter we discuss these types of experiments.

4.1.1 Laboratory experiments

Laboratory experiments are experiments that are characterized by its highly controlled nature, where the researcher can determine where, with whom and under what circumstances the experiment takes place. According to Charness and Kuhn (2011), strengths of the lab experiment include the ease of replicating results with newly acquired data, the ease of testing exact quantitative predictions from game theoretical models, and the ability to study behaviour where theoretical predictions are absent. Disadvantages are the short duration, which makes it harder to estimate long-term effects, the superficiality of the experimental context in general and the representativeness of the laboratory population in relation to the population of interest.

A specific form of a laboratory experiment is the survey. This type contains the beforementioned advantages of a lab experiment. However, Bauer, Ruof and Smeets (2018) state that surveys carry the risk of a biased response sample. This is especially relevant when studying a social theme, such as sustainable investing, as people with strong social preferences are more likely to participate in the experiment. Furthermore, in general people tend to give socially acceptable answers, while they do not put these answers into practice. This must be considered when determining a research experiment.

4.1.2 Field experiment

Field experiments are a type of experiment that are conducted in the everyday environment of the participants. The researcher is still able to manipulate the independent (treatment) variables, but not the exogenous ones. The treatment variable can be whether a child receives a computer from school, or whether a farmer receives a subsidy for more eco-friendly agriculture. Exogeneous variables are variables whose values are not determined within a model, for the quality of the teacher, or the level of rainfall. Advantages of field experiments are that the behaviour in a field experiment is more likely to reflect real life situations, and it is less likely that participants try to steer the experiment because they may not realize they are being studied[2]. A disadvantage is that there is less control over the exogenous variables that may incorrectly influence the result. An example of a field experiment in the setting of executive compensation models is the proposed in-company experiment.

4.1.3 Natural experiment

The natural experiment possesses similar characteristics as the field experiment, but now researchers no longer have control over the variable of interest. For example, a government policy enacted a regulation that changed the frequency of trash collection in one area, but not in the other area. The researchers did not have a say in it but can still study the effects of the policy. Advantages are similar to the field experiment, with an added advantage that a natural experiment can be applied in situations where it is not ethical to manipulate the independent variable. A limitation is that it is expensive and time consuming and similar to field experiments in the sense that biasing exogenous variables cannot be controlled.

4.2 Sample size calculations

When choosing an experimental design, it is important to determine the adequate sample size beforehand, because a larger sample size allows for a more accurate inference of the treatment effect. In this section we discuss the mechanism behind sample size calculations.

4.2.1 Basics of sample size calculations

The main objective of a sample size calculation is to determine the number of participants needed to detect a clinically relevant treatment effect (Noordzij, Tripepi, Dekker, Zoccali, Tanck & Jager, 2010). When working with different groups, e.g. due to stratification on certain participant characteristics, the sample size of each group must be large enough to detect significant effects. Hence, including multiple groups in an experiment requires the overall sample size to be bigger. This is relevant for our experiment, because when we are working with a treatment and control group, we have in the basis two groups. When we include risk aversion the number of groups increases to four, when adding personality traits, the number of groups increases even further. Therefore, sample size calculations are a relevant aspect of the design of the experiment to detect relevant effects.

The main advantage of sample size calculations is that interacting in the process of determining an appropriate sample size helps the researcher to formulate exact assumptions about the hypotheses to be tested (Gruener, 2020). In other words: the size of sample helps the formulation of the research hypotheses. Another advantage is that the objectivity of the research is strengthened by sample size calculations.

Whenever a researcher wants to perform sample size calculations, it should consider the following five parameters (Gruener, 2020):

- *Type I error*: the probability of rejecting the null hypothesis, while it holds true. It is usually denoted as the Greek letter alpha (α). The type I error could exist if a statistical test indicates differences between two groups, while such differences do not exist. This is known as the (usually 5 percent) significance level in other empirical methods.

---

[2] The idea that behaviour of the participant changes during the experiment when the person is aware of participating, is also known as the Hawthorne effect (Jones, 1992).

- *Type II error*: the probability of accepting the null hypothesis, while it is false in reality. It is usually denoted as the Greek letter beta (β). The type II error could exist if a statistical test does not indicate real differences between two groups, when there are in fact real differences.
- *Power*: the probability of correctly rejecting the null hypothesis, also known as the inverse of the type II error, i.e. $1-\beta$. The power is often set at 0.8. It is recommended to perform a power analysis before collecting data. In section 4.3 we further discuss the mechanism of power calculations.
- *Minimal relevant difference (MRD):* the minimum difference between treatment and control, also known as the effect size of the treatment. If the outcome variable is continuous, the MRD is a numerical difference. In the case of binary outcomes, event rates should be used. The MRD should also be economically relevant and in accordance with human behaviour in economic experiments.
- *Variability*: assumptions about the variability of the sample data are required and initially unknown, therefore they should be taken from pilot or similar studies. If the outcome measure is continuous, the variability of the sample data is the population standard deviation. In the case of a binary outcome measure, a formula of event rates should be used.

According to Gruener (2020), the result of the sample size calculation is influenced by the determination of parameters for the Type I and Type II error. Besides, it is also difficult to make assumptions about the MRD and variability. Sensitivity analyses can shed some light on which assumptions are systemically varied. In the case the data is not adequate and sample size calculations have to be conducted, researchers should refer to expert estimates, carry out pilot studies or at least argue about the standardized effect sizes.

An alternative to sample size calculations is the replication of experiments to aggregate evidence. However, according to Gruener (2020), replication cannot be a full replacement for sample size calculations because due to the costs, as respondents are typically paid for their participation. Experiments that are replicated are usually the ones that do not cost much and are easy to replicate, this is a problem for expensive experiments. Moreover, the incentives to replicate experiments are relatively low, as it is hard to publish replication studies (Galiani, Gertler and Romero , 2017).

When calculating the sample size, the researcher also has to account for binary or continuous outcomes. Then the sample sizes are calculated differently, which we discuss in the following two subsections.

4.2.2 Sample size calculations in case of continuous outcomes

According to Gruener (2020), the required sample *size N* for each group can be calculated according to the following formula.

$$N_{1,2}=\frac{2(\alpha+\beta)^2\sigma^2}{(\mu_1-\mu_2)^2}, \tag{1}$$

where α denotes the probability of a Type I error, β the probability of a Type II error, σ^2 is the variance of the population, μ_i is the population mean of group $i\in\{1,2\}$ and $(\mu_1-\mu_2)$ the treatment effect.

4.2.3 Sample size calculations in case of binary outcomes

According to Gruener (2020), the required sample size for each group can be calculated according to the following formula.

$$N_{1,2} = \frac{(\alpha+\beta)^2(f_1 \times \bar{f}_1 + f_2 \times \bar{f}_2)}{\mu^2}, \quad (2)$$

where α denotes the probability of a Type I error, β the probability of a Type II error, f_1 ($\bar{f}_1$) the fraction of people with (without) outcome 1 in treatment group 1, f_2 ($\bar{f}_2$) the fraction of people with (without) outcome 1 in treatment group 2 and μ is the treatment effect.

4.3 Statistical power

Recall that statistical power is the probability of rejecting the null hypothesis while it is indeed false in reality. In other words, the power can be seen as the probability to detect a significant effect when in reality there should be a statistically significant effect. Power calculations offer advantages such as being a tool for not over-interpreting findings of studies (Gruener, 2020). Additionally, power calculations help to answer the question of how much data should be collected (Vasilaky & Brock, 2020). However, it should be noted that power calculations are not related to causal inference or analysing data. Researchers should avoid power computations related to an observed effect, as power is not observable because it relates to ex ante concepts. Finally, power calculations have the advantage that the efficiency of the experimental design increases.

The research of Vasilaky and Brock (2020) offers a guide for researchers how to incorporate power calculations into their experimental research. They state that it is not always useful to report power calculations. In two scenarios reporting is justified:

1. Impossible to find a statistically significant effect due to the study being too underpowered.
2. For replicating studies and publishing adequately designed studies with null effects.

Consider the first scenario. When no statistically significant effect has been found after completing a study, an option is to calculate the minimum detectable effect (MDE) and variances. MDEs are calculations that help the researcher determining what the minimum impact an experiment should have to detect significant effects (Bloom, 1995). According to Vasilaky and Brock (2020), it is important to consider the confidence intervals around effects to report how accurate effect sizes are measured. When wide confidence intervals are reported, it can be a consequence of a low power or poor measurement.

Reporting under the second scenario is validated according to Vasilaky and Brock (2020) as it is an important part of the scientific process when researchers fail to detect an effect in a well-powered replication study, when others found significant effects. This result can be eligible for publication because it provides scientific evidence.

Low sample sizes make it difficult to calculate the power. Laboratory experiments often contain small sample sizes. Vasilaky and Brock (2020) discuss a few designs that exhibit small sample sizes and how to address power calculations:

- Multiple treatments: In the case of a between-subject design with multiple treatments a large sample size is required to detect a separate effect for each treatment on the outcome. The power calculation can be maximized by using an unbalanced design.
- Multiple hypothesis testing: When using multiple hypothesis tests the probability that at least one of them contains a type I error increases. This probability can be computed as:

$$1-(1-\alpha)^M, \quad (3)$$

Where M is the number of independent tests. When multiple testing takes place, the researcher can either adjust the power calculations before the experiment or after it by adjusting the Type I error rate.

Studies that possess similar populations and treatments can be used for determining the means of the control and treatment group (Vasilaky & Brock, 2020). In addition, these can also be applied to determine the variance of the population of interest. Special caution is needed when these are drawn from pilot studies of pre-intervention surveys because these often have small sample sizes. In the case of exploratory studies the power should be calculated over an array of effect sizes. The minimum detectable effect is useful to calculate when a specific expected effect is hard to calculate or when the sample size is limited.

4.4 Randomization

Another important methodological issue to consider is the randomization mechanism. To detect the treatment effect, the researcher has to properly address the issue of randomization. According to List, Sadoff & Wagner (2011), randomization ensures that the treatment is independent of other sources of variation is balanced across treatment and control group, making the estimate of the average treatment effect (ATE) unbiased. In this section some experimental designs as discussed in the research of List et al. (2011) are summarized. We follow their reasoning, unless otherwise specified.

4.4.1 Completely randomized design

In this experimental design treatments are probabilistically assigned to subjects independent of their observed or unobserved characteristics. The completely randomized design is the main design of our experiment. This design has the advantage of that the risk of treatment being correlated with individual characteristics is minimized. However, a weak point of the completely randomized design is that variances of outcomes could be potentially huge and sample sizes of treatment and control groups are randomly generated. This is even worse when dealing with a heterogenous subject group and the researcher is interested in decreasing the variance of the unobserved component. A solution is to include observable variables in the linear regression equation, in order to construct an estimate of the ATE with lower variance in finite samples.

4.4.2 Block design

Another solution for decreasing the variance when the subject group is heterogenous is to divide the experimental units into blocks, also known as blocking. As an example, a group of one hundred participants can be divided into five blocks that differ in total sizes. The intuition behind this technique is that heterogenous characteristics are handles as further treatments where randomization done is within blocks. This method has the advantage of allowing estimation of ATE's over subsamples of the subject group and it increases the efficiency of the experimental design.

4.4.3 Within-subject design

An alternative to the block design is the within-subject experimental design, where the same subject unit encounters more than one experimental treatment. In effect, the researcher blocks on a single subject. This design has the desired ability of increasing the accuracy of the ATE by decreasing the variance of the unobserved component. However, blocking on one subject may have the negative consequence of resulting in complicated interactions between treatments and hence lead to biased parameters. Another disadvantage is that treatments in a within-subject design may interact in ways that are not expected. An example is that participants may behave differently if they know what the best outcome is (learning effects).

4.4.4 Crossover designs

An extension of the within-subject design is the crossover design. Here the order in which treatments are applied to a subject is randomized. These can be used to solve the negative implication of blocking on one subject.

4.4.5 Factorial design

The factorial design is an approach that is useful for solving the problem of sample sizes that may vary across blocks in block designs. Efficiency of the experimental design is increased in factorial designs by choosing a pre-determined number of subjects to each combination of treatments. Randomization occurs over the order of treatments. This type of experimental design has the advantage of lower trial numbers. Less advantageous is that the factorial design leaves no space for controlling for interaction effects. In addition, basic factorial designs with equal sample sizes in each treatment block are likely to be inefficient.

4.4.6 Cluster design

Instead of allocating individuals to treatment groups, clusters of individuals are assigned within the cluster design. In this approach the outcome may occur at individual level, whereas randomization takes place at the cluster/group level. Hence, the unit of statistical analysis differs from the unit of randomization. Cluster design is useful because in field experiments there exist the possibility of correlation in the unobserved component among subjects within a cluster.

4.5 Measuring treatment effect

In this chapter we discuss various methods to measure treatment effects. We follow the reasoning of Harrison and List (2004).

4.5.1 Difference-in-difference regression

The difference-in-difference method (DID) can be used to estimate the treatment effect by comparing the difference in outcomes before and after for the treated group with the before and after outcomes for the control group. Key assumptions of the DID regression is that time invariant unit-specific shocks to the outcome variable that are correlated with the status of treatment and the assignment of treatment is independent of temporary individual-specific effect.

4.5.2 Propensity score matching

The objective of the propensity score matching (PSM) method is to perceive non-experimental data as experimental data. The intuition behind this method is that the treatment effect can be measured without bias, by determining observable factors such that two individuals with the same value for these factors exhibit homogeneous responses to treatment.

4.5.3 Instrumental variables

In order to measure treatment effects a variable can be found that is associated with treatment but is excluded from the outcome equation. These are so-called instrumental variables (IV). The method assumes that some components of the non-experimental data are random. However, less desired aspects of IVs is that these often do not exist or exist under undesired assumptions.

4.5.4 Structural approaches

Many structural approaches exist. These rely on complex estimation strategies with structural parameters. Structural approaches are common to use for ex ante policy simulation.

4.6 Incorporating methodological lessons in our experiment

In the first stage of our experiment, we perform controlled research using a Prolific panel of students and employees. We can classify this as a laboratory experiment. In a later stage we transform the laboratory experiment into a field experiment, namely an in-company experiment.

4.6.1 Laboratory experiment

In the laboratory experiment stage, we set up a survey. It has the advantage that surveys possess a highly controlled nature, where the researcher can determine where, with whom and under what circumstances the experiment takes place. However, we should be aware of biased results, due to that people with strong social preferences are more likely to participate in the experiment than people who do not have those preferences (Bauer et al., 2018). This can be solved by for example conducting lab experiments directly on the field population of interest, such that a more representative sample can be created (Charness and Kuhn, 2011).

Lab experiments also suffer often from having low sample sizes, which has adverse effects on power calculations. This can be solved by using multiple treatments or multiply hypothesis testing (Vasilaky and Brock, 2020).

It is important that the sample population resembles the target population, in this case CEOs. Using MBA or finance students, or employees with management experiences, accounts for this. Though these participants are not CEO, they are following a similar career path and some of them will eventually end up as CEO.

4.6.2 Sample size and power calculations

When data is not adequate/dependable, researchers should refer to expert estimates, carry out pilot studies or at least argue about the standardized effect sizes for the calculation of sample sizes (Gruener, 2020). For this purpose, the previous experiments can be used as reference material. Besides, when using an outcome variable that is either binary or continuous, different sample size equations should be used.

Reporting the power is only useful when (1) it is impossible to find a statistically significant effect when (due to an underpowered study) and (2) replicating studies and publishing adequately designed studies with null effects (Vasilaky and Brock, 2020). Confidence intervals for estimates should be reported because wide confidence intervals can be a consequence of low power or poor measurement.

Calculating the minimum detectable effect (MDE) and variances is useful when no statistical effect has been found after completing a study.

Literature

Ackerman, C. (2017, June 23). Big Five Personality Traits: The OCEAN Model Explained. PositivePsychology.com. https://positivepsychology.com/big-five-personality-theory

Agranov, M., & Tergiman, C. (2013). Incentives and compensation schemes: An experimental study. International Journal of Industrial Organization, 31(3), 238-247.

Andersen, D. J., & Lau, R. R. (2018). Pay rates and subject performance in social science experiments using crowdsourced online samples. Journal of Experimental Political Science, 5(3), 217-229.

Antia, M., Pantzalis, C., & Park, J. C. (2010). CEO decision horizon and firm performance: An empirical investigation. Journal of Corporate Finance, 16(3), 288-301.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. The Review of Economic Studies, 76(2), 451-469.

Bandiera, O., Barankay, I., & Rasul, I. (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *The Quarterly Journal of Economics*, *122*(2), 729-773.

Barker III, V. L., & Mueller, G. C. (2002). CEO characteristics and firm R&D spending. Management Science, 48(6), 782-801.

Bauer, R., Ruof, T., & Smeets, P. (2018). The effect of real and hypothetical choices on sustainable investment decisions in a large-scale field experiment.

Bauer, R., Ruof, T., & Smeets, P. (2021). Get real! Individuals prefer more sustainable investments. The Review of Financial Studies, 34(8), 3976-4043.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, *20*(3), 351-368.

Blanes, F., de Fuentes, C., & Porcuna, R. (2020). Executive remuneration determinants: New evidence from meta-analysis. Economic research-Ekonomska istraživanja, 33(1), 0-0.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation review*, *19*(5), 547-556.

Bolton, P., Mehran, H., & Shapiro, J. D. (2011). Executive compensation and risk taking. *FRB of New York Staff Report*, (456).

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: a meta-analysis. Journal of applied psychology, 89(5), 901.

Bruner, D., McKee, M., & Santore, R. (2008). Hand in the cookie jar: An experimental investigation of equity-based compensation and managerial fraud. Southern Economic Journal, 261-278.

Bugeja, M., Matolcsy, Z. P., & Spiropoulos, H. (2012). Is there a gender gap in CEO compensation?. Journal of Corporate Finance, 18(4), 849-859.

Bun, M. J., & Huberts, L. C. (2018). The impact of higher fixed pay and lower bonuses on productivity. *Journal of Labor Research*, *39*(1), 1-21.

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(1), 7-42.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab?. *Handbook of labor economics*, *4*, 229-330.

Cole, S., Kanz, M., & Klapper, L. (2015). Incentivizing calculated risk-taking: evidence from an experiment with commercial bank loan officers. The Journal of Finance, 70(2), 537-57

Daniel M. G. Raff, & Summers, L. H. (1987). Did Henry Ford Pay Efficiency Wages? Journal of Labor Economics, 5(4), S57–S86.

Eden, D. (2021). The science of leadership: A journey from survey research to field experimentation. The Leadership Quarterly, 32(3), 101472.

Eding, E., & Scholtens, B. (2017). Corporate social responsibility and shareholder proposals. *Corporate Social Responsibility and Environmental Management*, *24*(6), 648-660.

Edmans, A., Gabaix, X., & Jenter, D. (2017). Executive compensation: A survey of theory and evidence. The handbook of the economics of corporate governance, 1, 383-539.

Edmans, A., Gosling, T., & Jenter, D. (2021). CEO compensation: Evidence from the field. Working Paper. London Business School.

Edmans, Alex, and Qi Liu. "Inside debt." *Review of finance* 15.1 (2011): 75-102.

Elkinawy, S., & Stater, M. (2011). Gender differences in executive compensation: Variation with board gender composition and time. Journal of Economics and Business, 63(1), 23-45.

Gabaix, X., & Landier, A. (2008). Why has CEO pay increased so much?. *The Quarterly Journal of Economics*, *123*(1), 49-100.

Galiani, S., Gertler, P., & Romero, M. (2017). Open Science & Development Engineering: Evidence to Inform Improved Replication Models.

Garcés-Galdeano, L., & García-Olaverri, C. (2019). The hidden value of intangibles: do CEO characteristics matter?. International Journal of Manpower.

Gielen, A. C., Kerkhofs, M. J., & Van Ours, J. C. (2010). How performance related pay affects productivity and employment. *Journal of Population Economics*, *23*(1), 291-301.

Gilley, A., Dixon, P., & Gilley, J. W. (2008). Characteristics of leadership effectiveness: Implementing change and driving innovation in organizations. Human Resource Development Quarterly, 19(2), 153-169.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. American psychologist, 48(1), 26.

Gow, I. D., Kaplan, S. N., Larcker, D. F., & Zakolyukina, A. A. (2016). CEO personality and firm policies (No. w22435). National Bureau of Economic Research.

Grohol, J. M. (2019, May 30). The Big Five Personality Traits. PsychCentral. Retrieved 10 June 2020, from https://psychcentral.com/lib/the-big-five-personality-traits

Gruener, S. (2020). Sample size calculation in economic experiments. *Jahrbücher für Nationalökonomie und Statistik*, *240*(6), 791-823.

Ham, C., Seybert, N., & Wang, S. (2018). Narcissism is a bad sign: CEO signature size, investment, and performance. Review of Accounting Studies, 23(1), 234-264.

Harris, Q., Mercieca, A., Soane, E., & Tanaka, M. (2018). How do bonus cap and clawback affect risk and effort choice?: Insight from a lab experiment. Bank of England Staff Working Paper.

Harrison, G. W., & List, J. A. (2004). Field experiments. Journal of Economic literature, 42(4), 1009-1055.

Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. In Risk aversion in experiments. Emerald Group Publishing Limited.

Haynes, K. T., Josefy, M., & Hitt, M. A. (2015). Tipping point: Managers' self-interest, greed, and altruism. Journal of Leadership & Organizational Studies, 22(3), 265-279.

Holmstrom, B. (2017). Nobel Prize Lecture: Pay for performance and beyond. American Economic Review, 107(7), 1753-1777

Huang, S. K. (2013). The impact of CEO characteristics on corporate sustainable development. Corporate Social Responsibility and Environmental Management, 20(4), 234-244.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Research & Politics, 2(3), 2053168015604648.

Iqbal, S., Tian, H., Akhtar, S., & Sohu, J. (2019). Impacts of Performance-based Pay on Employee Productivity; Mediated by Employee Training. *International Journal of Research and Review*, *6*(10), 235-241.

Jones, S. R. (1992). Was there a Hawthorne effect?. *American Journal of sociology*, *98*(3), 451-468.

Jordan, C. E., Clark, S. J., & Waldron, M. A. (2007). Gender bias and compensation in the executive suite of the Fortune 100. Journal of Organizational Culture, Communications and Conflict, 11(1), 19.

Kajackaite, A., & Sliwka, D. (2020). Prosocial managers, employee motivation, and the creation of shareholder value. Journal of Economic Behaviour & Organization, 172, 217-235.

Kaplan, S. N., Klebanov, M. M., & Sorensen, M. (2012). Which CEO characteristics and abilities matter?. The Journal of Finance, 67(3), 973-1007.

Krause, R., Whitler, K. A., & Semadeni, M. (2014). Power to the principals! An experimental look at shareholder say-on-pay voting. Academy of Management Journal, 57(1), 94-115.

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

Lazear, E. P. (2018). Compensation and Incentives in the Workplace. Journal of Economic Perspectives, 32(3), 195-214.

Lim (2020). Big Fiver Personality treats. Link: Big Five Personality Traits | Simply Psychology. Accessed on 18-10-2021

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439-457.

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. The journal of finance, 60(6), 2661-2700.

Malmendier, U., & Tate, G. (2015). Behavioural CEOs: The role of managerial overconfidence. Journal of Economic Perspectives, 29(4), 37-60.

Mirrlees, J.A. (1999). The theory of moral hazard and unobservable behaviour: Part I. Review of

Nadkarni, S., & Herrmann, P. O. L. (2010). CEO personality, strategic flexibility, and firm performance: The case of the Indian business process outsourcing industry. Academy of Management Journal, 53(5), 1050-1073

Noordzij, M., Tripepi, G., Dekker, F. W., Zoccali, C., Tanck, M. W., & Jager, K. J. (2010). Sample size calculations: basic principles and common pitfalls. Nephrology dialysis transplantation, 25(5), 1388-1393.

O'Reilly III, C. A., Caldwell, D. F., Chatman, J. A., & Doerr, B. (2014). The promise and problems of organizational culture: CEO personality, culture, and firm performance. Group & Organization Management, 39(6), 595-625.

Ottersen, S. S. (2016). *Does a Change in Performance Pay have an Effect on Workers' Productivity?-An empirical analysis of a change in performance pay in Company X* (Master's thesis, University of Stavanger, Norway).

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioural and Experimental Finance*, *17*, 22-27.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411-419.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioural research. *Journal of Experimental Social Psychology*, *70*, 153-163.

Resick, C. J., Whitman, D. S., Weingarden, S. M., & Hiller, N. J. (2009). The bright-side and the dark-side of CEO personality: examining core self-evaluations, narcissism, transformational leadership, and strategic influence. Journal of Applied Psychology, 94(6), 1365.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. Journal of research in personality, 43(2), 137-145.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological science, 2(4), 313-345.

Rosen, B., & Jerdee, T. H. (1976). The influence of age stereotypes on managerial decisions. *Journal of applied psychology*, *61*(4), 428.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *The Review of Economic Studies*, *71*(2), 513-534.

Vasilaky, K. N., & Brock, J. M. (2020). Power (ful) guidelines for experimental economists. *Journal of the Economic Science Association*, *6*(2), 189-212.

Weel, B. ter, Witteman, J., Verheuvel, N. (2019). *Executive compensation and firm performance.* SEO Report Nr. 2019-30, Amsterdam: SEO Amsterdam Economics.