Author

I. Djordjevski

Crowd versus Experts Forecasting Technologies

Impact of Collective Diversity & Size on Collective Performance

October, 2021

DELFT UNIVERSITY OF TECHNOLOGY

MASTER THESIS

Crowd versus Experts Forecasting Technologies

Impact of Collective Diversity & Size on Collective Performance

Author:
I. (Igor) Djordjevski

Student ID:
4243722

Graduation Committee:

Chair	**Dr. G. (Geerten) van de Kaa**
First Supervisor	**Dr. G. (Geerten) van de Kaa**
Second Supervisor	**Dr. J. (Jafar) Rezaei**

A thesis submitted in fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

MANAGEMENT OF TECHNOLOGY

Faculty of Technology, Policy and Management

To be defended in public on November 3rd 2021

TU Delft Delft University of Technology

Declaration of Authorship

I, I. (Igor) Djordjevski, declare that this thesis titled “Crowd versus Experts Forecasting Technologies – Impact of Collective Diversity & Size on Collective Performance” and the work presented in it are my own. I confirm that:

- ❖ This work was done wholly or mainly while in candidature for a research degree at this university.
- ❖ Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- ❖ Where I have consulted the published work of others, this is always clearly attributed.
- ❖ Where I have quoted from the work of others, the source is always given. Except for such quotations, this thesis is entirely my own work.
- ❖ I have acknowledged all main sources of help.
- ❖ Where the thesis is based on work done by myself jointly with others, I have clarified what others did and what I have contributed myself.

Date: ____________________________

Signature: ____________________________

“Two heads are better than one, not because either is infallible, but because they are unlikely to go wrong in the same direction.”

– C.S. Lewis

Delft University of Technology

Executive Summary

Crowd versus Experts Forecasting Technologies

Impact of Collective Diversity & Size on Collective Performance

By I. (Igor) Djordjevski

For centuries, Homo sapiens have been trying to predict the future through supernatural or scientific methods. Since prehistory, this quality has been essential to humans (e.g., anticipate prey, then ambush it). Even in today's society, predicting the future remains essential in many industries and research domains. Although, we are still far from producing flawless forecasts (e.g., weather) because future events are uncertain. When decisions contain uncertainty, governments, organizations, and individuals alike tend to be interested in the advice of others.

One such case that was interesting is when predicting the outcome of a standard battle. Within such a battle, high-tech firms compete to obtain most customers in a given market through their technological inventions. To predict which technology will be the standard on the market, experts are independently interviewed to determine the importance of factors (e.g., weights) that can influence this battle. Were the second round of expert interviews need to assign a value for each of these factors for every competing technology—resulting in a performance grade used to make the prediction. This prediction indicates which firm/technology will likely have the upper hand in the market. The factors originate from the list of factors in combination with the Best-Worst Method (BWM), which allows evaluating the multiple-criteria decision-making (MCDM) problem (Rezaei, 2015a; v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011).

Multiple studies expressed concerns that it is challenging to persuade and find experts willing to participate in the interview. Instead of finding a better way to approach the experts, this study focused on another solution not applied before in predicting standard battles. Hence, the objective was to understand, test, and examine how the Collective Intelligence (CI) of the crowd (i.e., group of random individuals) performs compared to experts. The idea of CI is that it does not reside in any individual but emerges from the group. When people's opinions are combined, their advice should be more truthful or similar to a typical expert's.

In other words, this quantitative exploratory study investigated if CI in comparison to experts differs when predicting standard battles. Hence, a literature review was required to provide deeper insights and factors that influence CI. This study explored the underlying mechanism of CI and established a conceptual model based on the theoretical background, which indicates the (moderating) relationship between 'Diversity' (DIV), 'Group Size' (GS), 'Performance' (PERF) of the crowd.

The variable DIV was measured based on differences in gender, age, degree, job, and nationality and expressed by the Simpson's index, reflecting the number of different species and distributions (SIMPSON, 1949). As for GS, the only attribute measured was the number of people in a contrived group. Further, the definition of the PERF of a collective is the quantifiable difference in their solution relative to the prediction proposed by the experts (Wagner et al., 2010). Hence, this was dubbed 'Relative Performance' (RP) for the rest of this paper.

Prior research on standard battles was selected to test and validate the assertions in this study. This selection was based on several factors, such as the outcome of the battle was predicted by experts but where the ground truth is unknown. Doing so allowed validating if the crowd performs differently than experts in prediction tasks. The selected study involved the battle of two wind turbine technologies (WTT) dubbed 'Gearbox' (GB) and 'Direct Drive' (DD) (van de Kaa et al., 2020). Further, to interview the crowd, the traditional BWM questionnaire was converted into a cross-sectionally distributed online survey (i.e., MTurk) to obtain the data. Two hundred respondents completed the survey, but only 137 remained after pre-qualifications.

In this research, 'groups' were contrived from the sample that completed the prediction task. Here, the group members employ their expertise to carry out the given task. In other words, all group members performed the same activity and then were randomly pooled based on varying group sizes. Next, simplified random sampling was carried out twice. The first time this resulted in six groups with a respective size of 5, 10, 20, 30, and 40. A second time was required because the former did not provide a good range of DIV scores. Because of this limitation, the sample was sampled multiple times, which resulted in multiple subgroups with each a DIV score—resulting in five groups of size 5, 10, 20, 30, 40, and 137 with respectively 26, 13, 6, 4, 3, and 1 subgroup for comparison.

After sampling the data, the variables were tested for normality and homoscedasticity. The results indicate that the independent variables (i.e., DIV, GS) do not satisfy the normality assumption. Hence, the variables are not suited for parametrical testing. However, both the non- and parametrical tests were applied. Because ANOVA seems not to be very susceptible to modest divergence from normality. Namely, various studies used a variation of non-normal distributions and concluded that the false positive rate is not affected much if the notion of normality is not satisfied (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). In addition, it was required to compare more than two groups. Hence, the One-Way ANOVA test (OWAT) and the Kruskal-Wallis Test (KWT) were selected to investigate the relationship between the variables (DIV, GS, RP). More specifically, the effects that GS or DIV can have on the RP of the crowd. In addition, due to systematic limitations with simplified random sampling, Bootstrapping (BOOT) and Monte Carlo (MC) were also performed respectively for the OWAT and KWT to investigate the relationship between DIV and RP. Lastly, the moderation effect was tested based on the Linear Regression (LR) method.

The results did not show any significant differences between DIV and RP, and any significant moderation effect. Hence, the initial hypothesis that a more diverse group of individuals would perform better was refuted. In addition, this research rejects the proposition that the relationship between GS and RP should be positively moderated by how diverse the crowd is. Consequently, this research was not able to conclude how these variables affected the PERF of the crowd. Nevertheless, the former results weaken the theory of (S. Krause et al., 2011; Nguyen et al., 2018; Surowiecki, 2005), who underlines the importance of DIV. In contrast, the results gave ample support to (Reynolds et al., 2017) and their claim that there is no correlation between DIV and PERF. However, their second assertion about an existing relationship between cognitive DIV was not tested. Hence, it is recommended that future studies investigate this relationship.

The findings did show a significant difference between groups for the variable GS in the case of GB WTT. Namely, when the size of the group increases, the PERF also proportionately increases with an upper limit, dubbed 'Optimal Group Size' (OGS). A U-shape relationship defines this relation between the variables. However, both the OWAT and KWT provide contradicting findings. Namely, findings from the OWAT suggest that the OGS consists of 15 people and that there is indeed a U-shape relationship between GS and RP, which proves that the claim of (Hashmi, 2005) and our hypothesis is correct. In contrast, the KWT indicates a relatively linear relationship, where the group of 10 and 20 performed significantly better than 30. However, OWAT showed that the group of 30 performed significantly better than the group of 10 and 20 people. Whereas the OGS was 10 and not 15 individuals. The results from the (non) parametrical for the location of OGS indicate that a group of 10 or 15 outperforms the other groups of smaller and bigger sizes. Hence, supporting the claim of (Carvalho et al., 2016) and weakening (S. Krause et al., 2011).

To conclude, how CI operated in this research depended on survey completion time, consistency ratio, selection of best and worst criteria, PERF grade, and the final prediction. Hence, this research concluded that the CI of the crowd did show differences in predicting the outcome of a standard battle compared to the expert pool. This was primarily based on the fact that the crowd had a PERF score that was two-thirds lower than that of the experts. Although, the crowd performed in some aspects in similar or better ways. This research only tested one case, limiting our insights if this happened due to chance or not. In addition, the main goal of this study was to investigate if the crowd could come up with the exact prediction. Meaning that the conclusion and the process (e.g., selection criteria and grading WTT) towards this conclusion should be similar to experts. However, this was not the case. In addition, the consistency ratio remains a matter of doubt since most of the results were initially unreliable, which required (logical) corrections to obtain reliable results. Making the comparison based on the consistency ratio rickety. To reiterate, the crowd performed differently than experts when predicting the outcome of the WTT battle. Despite various differences, the individual and groups would conclude a similar prediction like in the case of experts.

Preface

Dear Reader,

Before you lies the conclusion of my educational time and the dissertation "Crowd versus Experts Forecasting Technologies – Impact of Collective Diversity & Size on Collective Performance", the basis of which is an online survey on Collective Intelligence conducted amongst several random individuals. Furthermore, it fulfills the Master of Science (MSc) in Management of Technology at the Delft University of Technology. I was engaged in researching, gathering respondents, and writing this dissertation from February to October 2021.

Before this moment, I used to consider myself a slacker when I was in High School. Heading into college, my lazy ways continued. Usually, when a narrative starts in this way, it doesn't have a good ending. Fortunately for me, my 'hunger' for knowledge started growing when I was in my second year of uni. Since then, the hunger has exceeded my wildest expectations, bringing me mysterious new food for my never-ending thoughts and stomach every day. Although the integration of academic learning was supported and provided by the university, my personal improvement and individual growth were not. Thus, I needed to take responsibility for my learning and development. This view molded a picture about me as a sculptor whose body is his unique tool for life. To make proper use of my tool, I sculpted my centered and holistic experience by focussing on understanding values, nurturing skills, and moving towards cross-disciplinary knowledge while being reflective about what I am doing and where I am heading. From this point of view, my final work became a measure for me personally to see what fruits I could reap from the hard labor I put into myself all these years. Seeing these kinds of results only makes me more self-satisfied.

However, the motivation that led me to this moment did not entirely come out of the blue. Instead, others that I met during my lifetime influenced this to some degree. Probably, to some extent, I am the person today in some part because of these individuals. Therefore, I want to say this to Mr. E. Centen (Elementary school teacher), dhr. L. Kempe (Mentor & Middle school biology teacher), and the current Vice-Rector Magnificus / Vice President Education (Prof. R. F. Mudde) of the TUDelft, thank you for caring for me and helping me understand my value to myself, as well as realizing that I am capable of more than I could ever think of.

During the previous six months, I have been working to understand the use of Collective Intelligence for complex tasks, such as predicting standard battles. This subject intrigued me because it was a phenomenon that was unknown territory for me. Combined with something I already was interested (standard battles) in, resulting in a subject worth exploring because interdisciplinary knowledge and thinking were involved.

Even though most of us were locked indoors due to the pandemic, I received guidance through digital means thanks to video calls. Although this was not without any hurdles, it made it possible for me to master this project from the comfort of my own home without ever meeting the supervisors in person. Hence, making this achievement even more special. But this would not have been possible without my graduation committee members, who provided their input and suggestions during this project. I appreciated that they allowed me to have total freedom and autonomy to investigate an intriguing subject. I would like to thank my graduate committee, Geerten van de Kaa, for his creativity and excellent discussions during all the project stages. Also, your enthusiasm provided me with ample motivation to continue this journey. Second to Jafar Rezeai, thank you for your support and invaluable advice and highlighting the areas contributing to the project.

Due to the COVID-19 pandemic, I could not be physically present at the university for more than a whole year, making it more challenging and very educational in a personal and professional manner to graduate. I will cherish the many moments I had as a student. For all the beautiful individuals in this world, if you can dream it, you can do it! Thank you to everyone that has offered their help and support in making this experience better than expected.

In addition, to the ending of this chapter of my life, where I developed both personally and academically, I could not have made this possible without dedication, willpower, and a twist of stubbornness, so this part is for you. Thank you.

– Igor Djordjevski –

Contents

List of Figures

List of Tables

List of Operational Definitions

	Abbreviation	Definition
A	*AHP*	*Analytical Hierarchy Process*
	ANP	*Analytic Network Process*
B	*BC*	*Bonferroni Corrections*
	BWM	*Best-Worst Method*
	BOOT	*Bootstrapping (SPSS)*
C	*CI*	*Collective Intelligence*
	CS	*Crowdsourcing*
D	*DD*	*Direct drive*
	DIV	*Diversity*
	DV	*Dependent Variable*
G	*GB*	*Gearbox*
	GQ	*Gold question*
	GS	*Group size*
H	*HREC*	*Human Research and Ethics Committee*
I	*IV*	*Independent Variable*
K	*KST*	*Kolmogorov-Smirnov Test*
	KWT	*Kruskal-Wallis Test*
L	*Ls*	*Levene's statistic*
	LSD	*Least Significant Difference*
M	*MCDM*	*Multiple-Criteria Decision-Making*
	MCQ	*Multiple-choice questions*
	Md	*Mahalanobis Distance*
	MC	*Monte Carlo (SPSS)*
O	*OGS*	*Optimal Group Size*
	OWAT	*One-Way ANOVA Test*
P	*PCS*	*Pearson Chi-Square*
	PERF	*Performance*
R	*RP*	*Relative Performance*
S	*SI*	*Swarm Intelligence*
	SWT	*Shapiro-Wilk Test*
W	*WoC*	*Wisdom of Crowds*
	WTT	*Wind turbine technology*

To

Jovanka & Ljupco Djordjevski

For providing me with this opportunity

Chapter 1

Introduction to the Study

Chapter overview

This introductory chapter describes the motivation of this study and the approach to answering the identified research problem, objectives and the value of possible outcomes are highlighted and described.

➢ *The structure of this chapter is as follows: Section 1.1, begins with a general introduction of the phenomenon, such as similar other concepts, history, research, and findings, selected definition. Section 1.2 provides general background information that connects the concepts of standards battles and the Best Worst Method (BWM) to Collective Intelligence in this research. Section 1.3 discusses the current issues with standard battles and BWM literature, for which this research aims to find a solution. Subsequently, Section 1.4 explains the research objective, questions, and deliverables of this research. Lastly, the general structure of this thesis is described and visualized in Section1.5.*

1.1 Introduction

Some may look into the stars to predict the future; some may use past performances to foresee the potential outcome through statistical means. When governments, organizations, and individuals alike ought to make choices under uncertainty, they have a tendency to be interested in the opinion of others. To obtain new information that can inform their own choices. In particular social sciences, researchers have shown that the voice of knowledgeable individuals influences individuals (e.g., experts) (Felton et al., 2013; Milgram, 1963; Tyler et al., 1992). Undeniably, we all possess some form of intelligence, some more than others. However, intelligence can also appear in groups, where individuals act collectively to combine their knowledge and insight. Also known as CI, and does not reside in any individual but emerges from the group. Other scholars observed this phenomenon and focused on studying the advice generated through crowdsourcing (CS) (Bhatti et al., 2020). Currently, many similar concepts are researched, such as Open Innovation, User Innovation, Co-creation, Open Source, and more (Benkler, 2006; Chesbrough, 2003; Thomas W Malone et al., 2003; Tapscott et al., 2008; von Hippel, 1986) Nevertheless, this research will only focus on CS to leverages the CI of a group of individuals. `

The idea for CI resides on the principle that everyone has knowledge that is valuable to other people. When a group of independent individuals and their advice is statistically aggregated (e.g., measures of central tendency), it will be more truthful when compared to the mean of a typical expert due to exploiting the benefit of error cancellation. In other words, diversity can compensate for the bias of a small group, for instance, with a relatively small number of experts. However, how can it be that with this seemingly-too-naïve-to-be-real approach, someone with no experience nor knowledge about the subject in question can decide on what to do? Many researchers attempted to explain this phenomenon and its mechanisms but without scientific consensus. Nevertheless, an attempt is made with the following example. There is a group of different individuals. Each person has their way of thinking and has different kinds and amounts of experience and knowledge. These differences are all then gathered to aggregate the results on a particular subject, resulting in an answer more accurate than experts. Sadly, throughout history, this point of view was not always shared. Around the 18th and 19th century the elites of society viewed the crowd as problem generators (Wexler, 2011). After reconfiguring this concept in the social sciences, the view shifted from a problem generator to a problem-solver and innovator (Wexler, 2011), which probably materialized due to various research demonstrating that aggregated crowd advice can perform better (i.e., decision performance and accuracy) than informed individuals or experts (Adomavicius et al., 2005; Budescu et al., 2015; Larrick et al., 2012; Lorenz et al., 2011). Others showed crowd-based suggestions on various predictions tasks to be as accurate as knowledgeable individuals or better (Larrick et al., 2012; Mollick et al., 2016; Ray, 2006; Sunstein, 2008; Tetlock, 2017). While additional researchers proved that advice from the CI reduces the variance of recommendations to its mean, enabling superior decision-making compared to that of an individual (Chiu et al., 2014; Kittur et al., 2008; Kozinets et al., 2008; Lorenz et al., 2011). One would probably suppose that this phenomenon would only work in a setting with yes or no questions. However, this is not the case, as Surowiecki showed by providing several examples in his book. He concluded that group-based decisions also work for matters where solutions are more complicated than binary answers (Surowiecki, 2005). In addition, the Internet can be of great use to leverage the advice from the CI of a group of individuals due to enhancements in efficiently disseminating information (Chiu et al., 2014; Woolley et al., 2015). Harvesting such individual wisdom has been applied for several purposes, such as learning, leveraging skills, recurring cost and processing time, problem-solving (i.e., wicked problems), (managerial) decision-making and predictions, generating knowledge, and capital investment (Allahbakhsh et al., 2013; Bhatti et al., 2020; Budescu et al., 2015; Chiu et al., 2014; Howe, 2006; Larrick et al., 2012; Levy et al., 1997; Thomas W. Malone et al., 2009; McGrath, 1984; Mollick et al., 2016; Sloane, 2011; Suran et al., 2020; Zhao et al., 2014). In conclusion, when considering the characteristics of a standard battle and of experts as mentioned previously, one can conclude that characteristics of CS and CI are similar, and thus, a suitable candidate for this research to investigate as an alternative solution.

The research area for CS and CI has been active in the past couple of years. By offering a wide variety of research fields ranging from economics, biology up to information systems and many others (Bhatti et al., 2020; T.W. Malone et al., 2015; Suran et al., 2020). This broad spectrum can also be found in the industries leveraging the notions, such as health care, gaming, education, transport, public transport, etcetera (Bhatti et al., 2020). This increase in attention is to some extent because the concepts are relatively simple to understand and utilize, but primarily due to the vast invasion of Internet technologies and smartphones (Nalmpantis et al., 2019), making it easier to access the crowd on a global level. Several other reasons for organizations to use CS are limited resources (e.g., time, money) and capabilities (e.g., in-house expertise), focus on core competency, and obtaining answers to wicked problems (Chiu et al., 2014; Schenk et al., 2009).

In 2006, Howe dubbed the term CS, arguing that companies employed the Internet to reach respondents for a task (Howe, 2006). Nevertheless, this concept was not new. Namely, at the end of the 1970s, CI was introduced, and after some years, the concept got validated in the 1990s (Levy et al., 1997; T.W. Malone et al., 2008). Still, Howe inspired various researchers alike. He based his work on User Innovation (von Hippel, 1986) along with others. In 2010, Jeff redefined CS as "The act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call." (Vander Schee, 2009). Other researchers and practitioners that studied CS and CI, offer several definitions with subtle differences. When combined, these views define CS and CI as a technique for solving a particular assignment by collecting dispersed individual skills in an online or offline context (Brabham, 2008; Estellés-Arolas et al., 2012; Levy et al., 1997; Lorenz et al., 2011; Thomas W. Malone et al., 2009). Furthermore, the notion CS is synonymous with CI. Even though CS is an umbrella term for types such as CI (Chiu et al., 2014), the difference between the two ideas is minor, and some use these notions interchangeably. Shahzad Sarwar and others found numerous taxonomies varying from mass collaboration to crowd wisdom and others (Bhatti et al., 2020), which could explain the sometimes blurry and ambiguous definitions found in the literature.

As mentioned before, many definitions and similar concepts are studied; incorporating various perspectives has yet to materialize and may be unfeasible (Bhatti et al., 2020). As a result, the difference between the two concepts should be clear and easy to understand. Therefore, for the remainder of this paper, CS aims to distribute the workload from one to many. On the other hand, CI is a goal, where the focus lies on the output. From this point of view, the generic term "CS," for which an explanation is used is similar to "CI." Thus, the following definition for CS is used because it considers various characteristics such as task creator, group members, processes and includes the use of a crowd for problem-solving:

> ***"Crowdsourcing is an online distributed problem-solving paradigm, in which an individual, company, or organization publishes defined task(s) to the dynamic crowd through a flexible open call to leverage human intelligence, knowledge, skill, work, and experience." (Bhatti et al., 2020)***

To conclude, a crowd can perform better at a job than small groups, individuals, and even experts. The crowd exists out of individuals from all over the world, with various backgrounds. Individuals can make their assessments on the same topic independently. When combining these decisions, the crowd's judgment would be incredibly close to the truth. The aggregate judgment is often more accurate, even when predicting the future, rendering it a source of intelligence, which can be exploited and used for various purposes. Hence, this research applies CI in the context of predicting standard battles to investigate assertions from different studies. However, for one purpose, no research has applied CI in the context of predicting standard battles. Hence, this research investigates if the previous assertions are true. The following section provides background information about standard battles and the Best-Worst Method (BWM).

1.2 Background Information: Standard Battles

Nowadays, technological breakthroughs are increasing in numbers and their impact, as they reshape our future, affecting our lives and those of businesses. Specifically, technologies from high-tech companies are replaced more quickly by the technology of other competing firms. However, these technologies also increase the number of companies who market their product (standard) simultaneously on the market, creating even more competition. This form of competition is called a 'standards battle'. Suppose a firm's technology becomes the standard (e.g., market leader). In that case, their position allows them to gain control and collect considerable monetary rewards. In various studies, a firm's standard becomes dominant when that company holds more than 50% of new customers from that market for a considerable amount of time (Suarez, 2004). There are three distinct mechanisms (i.e., market-, committee-, and government-based standardization) to predict the battle's winner (Wiegmann, P.M., de Vries, H.J., Blind, 2017). Mixing two or more methods is dubbed multi-mode standardization.

Several scholars have explored the prediction of 'ante' or 'post' battle. Their work provided more insights into standard battles by developing frameworks. Some established and frequent applied works are (Lee et al., 1995), (Schilling, 1998), (Suarez, 2004), and (v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011). These frameworks include factors for standard dominance. These are imperative to predict the likelihood of a standard becoming dominant because factors can positively or negatively affect this likelihood. As for the focus of this research, the work of v.d. Kaa is of interest because the framework was the most comprehensive thus far (van de Kaa et al., 2015). The selected framework has a total of 29 factors that can affect the chance of a standard becoming dominant (v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011).

This is only the first component that is required to make the prediction. The second part is a frequently applied methodology in combination with the works of Geerten. This technique was developed by Jafar and was dubbed the Best Worst Method (BWM) (Rezaei, 2015a). This method allows evaluating the Multiple-criteria decision-making (MCDM) problem and selecting a dominant standard. The BWM utilizes the responses from experts' to establish weights for various factors. However, in the literature, other methods were also used to solve similar problems, such as Analytical Hierarchy Process (AHP), Analytic Network Process (ANP), TOPSIS, ELECTRE, PROMETHEE, (Rezaei, 2015a; Zanakis et al., 1998). From these options, AHP and BWM were applied frequently in several research studies investigating standards battles. Nevertheless, the BWM has some advantages over the AHP method. Namely, the BWM needs fewer comparison data than AHP while also providing a consistency ratio (Rezaei, 2015a). Hence, this study is only interested in the BWM. Lastly, when applying the BWM, a minimum of five interviewees is required to determine the factor's weights and technologies ratings (v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011).

1.3 Research Problem

To understand and comprehend predictions for standard battles, additional explanations are provided. These predictions are done with the work of Geerten and Jafar and are synthesized as follows. The selected framework has a total of 29 factors sorted into five categories (i.e., (1) Characteristics of format supporter, (2) Characteristics of format, (3) Format support strategy, (4) Other stakeholders, (5) Market characteristics) (v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011). Here, a factor is relevant if mentioned in the literature or by two or more experts who mentioned the factor being relevant during an interview—resulting in a list of relevant factors that can influence the prediction.

The next step of the prediction process is to calculate the weights of the important factors by interviewing additional experts from a particular industry. By doing so, more significant insights into a phenomenon are realized (Massam, 1988). The interviewees first receive an explanation of definitions, factors, categories, and the process of the BWM. Subsequently, they are independently questioned to decide the worst and best factors from the list of factors presented to them. Next, they compare the best and then their worst factors with the factors from the list by grading them. Also known as a pairwise comparison, resulting in a maximin problem which is solved to obtain the significance of all the factors.

The next step is to evaluate or grade how the technologies perform based on these factors by interviewing the second group of experts. Based on a seven-point Likert scale, the experts assign values to the factors for each technology, which is vital because it shows which standard is likely to dominate the market. These valuations are then averaged and multiplied with the weights of the factor determined by the first round of expert interviews. Resulting in the final performance grade per technology. Finally, this grade is used to predict which technology will be dominant in the market.

Unfortunately, using only experts for the prediction contains some constraints. Therefore, this research focuses on finding a suitable solution for a (particular) drawback. Namely, one of the main bottlenecks of conducting experts interviews for this method is that relatively small numbers can raise issues of representativity and specificity. Even though there are plenty of experts, these are not always easy to identify, let alone convince these individuals to spend their time and energy on an interview that takes roughly two or three hours without receiving compensation. This limitation can be seen in multiple studies, especially in standard battles and the use of the BWM. Here, researchers often state that it is challenging to find experts with adequate knowledge and experience to partake (Anantapantula, 2017; Ram, 2018; Tedjakusuma, 2014; Tehrani, 2014; van de Kaa et al., 2017; Zaarour, 2011), which is troublesome because relatively small numbers can raise issues of representativity and specificity or may even introduce biases depending on the statistical method used (e.g., convenience sampling) (Ram, 2018). In addition, the custom is to state that future studies should conduct more interviews with experts (van de Kaa et al., 2019). However, this is contradicting because if finding experts with the right expertise is unsuccessful, then the solution proposed to find even more experts in future research will probably yield nothing. Therefore, the following research problem emerges.

Problem Statement: ***The paucity of experts to conduct the BWM questionnaires in the context of standard battles negatively influences the representativity and specificity of the experts' predictions.***

It is crucial to understand that conducting interviews with experts can be considered a type of CS. The potential solution is a similar CS solution, the 'Collective Intelligence.' Tracing back this resolution leads to Aristotle's theory of collective judgment, as presented in his work Politics, by combining multiple opinions or forecasts from groups of random individuals. He noticed that when individuals see distinct parts of the whole, the collective appraisal can surpass (knowledgeable) individuals or derive a consensus from interactions between the experts in the group (Sunstein, 2008). In addition, (Hong et al., 2012) confirmed Aristoteles' view on the existence of CI by observing the democratic process. Lastly, in the literature review, the explanation of this phenomenon is further discussed and elaborated.

1.4 Research Objectives & Questions

Based on the above, this research aims to address the problem identified in Section 1.3. To summarize: an issue hampering the representativity and specificity of expert predictions is the lack of available, knowledgeable, and inclined experts to participate in the BWM questionnaires.

In addition, the custom is to state that future research should continue the research and interview more experts. However, this is contradictory because finding experts with the right expertise is often unsuccessful. The proposed solution of finding more experts for future research will therefore probably yield nothing. From this point of view, this research will focus on investigating another solution. Therefore, the following overarching objective of this study materializes.

Research Objective: ***Understand, test, and examine how the Collective Intelligence of a crowd performs compared to experts.***

In acknowledgment of the above, the main research question guiding this research is:

Main Research Question

How does the Collective Intelligence perform in comparison to experts when predicting technologies?

To solve the main research question, a total of five corresponding sub-questions are derived:

Sub-question 1: ***Which factors influence the Collective Intelligence of a crowd when performing a task?***

Sub-question 2: ***How does group size affect collective performance?***

Sub-question 3: ***How does group diversity affect collective performance?***

Sub-question 4: ***How does the moderation effect of diversity affect collective performance?***

Sub-question 5: ***What do the results from the Collective Intelligence of the crowd indicate in comparison to the prediction done by experts?***

The literature review provides the required theoretical background to help answer the first sub-question. Three deliverables were created based on the literature review: Exploring the under-researched topic of Collective Intelligence (CI) in the context of Multi-Criteria Decision-Making (MCDM) problems, systematically describing the characteristics affecting CI, and establishing the relationship between diversity, size, and crowd performance.

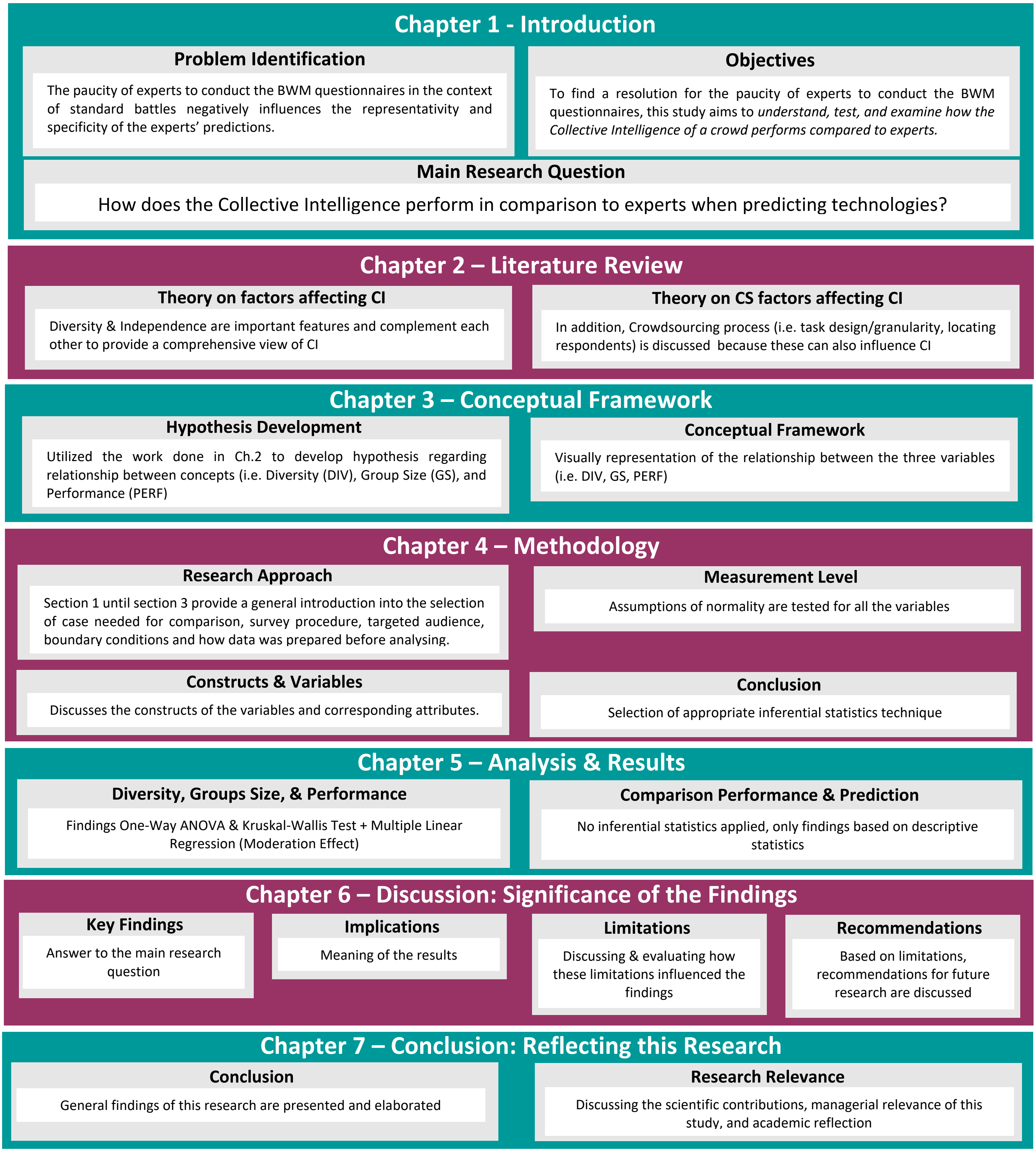
To test the conceptual model, which answers sub-questions two, three, and four required a quantitative method to examine and test the hypothesis that diversity and size affect the performance of a crowd. By doing so, it allowed to explore the association between size, diversity, and performance. This was achieved by combining secondary data (i.e., prior research on standard battles) and primary data. The latter entails descriptive data obtained by converting the traditional BWM into an online distributed survey. These observations occurred without the intervention of the researcher.

One prior research on standard battles was selected to acquire the required secondary data—namely, the study involving the battle of wind turbine technology (WTT). More specifically, the gearbox (GB) versus direct drive (DD) technology (van de Kaa et al., 2020). This decision was based on several factors, such as that experts predicted the battle's outcome, but the ground truth of this prediction was not verified. Doing so allowed validating if the crowd performs differently than experts in prediction tasks. Other aspects that influenced this choice were the availability of data of the prior study, but also time and monetary expenses as these resources were limited.

1.5 Thesis Structure Guide

To achieve the objective of this research study, this dissertation follows the typical structure for scientific research. Figure 1.1 shows the complete structure of this thesis and specific information.

This chapter, Chapter 1, has specified the research problem, objectives, and questions to be answered through this research study. The following chapter, Chapter 2, reviews relevant literature and provides the required theoretical background for the questions formulated in Chapter 1. Furthermore, more insight into CI and the different factors that can affect CI are discussed. In Chapter 3, the theoretical background of Chapter 2 will ground this project, leading to a conceptual model showing the (moderating) relationship between diversity, size, and performance. In Chapter 4, the research approach and methodology are further elaborated, including gathering empirical data and analyzing the variables to find a suitable inferential statistic. Subsequently, Chapter 5 presents the results from two inferential statistical tests (One-Way ANOVA and Kruskal-Wallis) and discusses their meaning. Chapter 6 is the chapter that presents the answers to the main research question. Furthermore, the interpretations, implications, limitations, and recommendations are discussed. Finally, in Chapter 7, this research's overall conclusion is elaborated and discusses the relevance of this research for the scientific community, societal and practical-, and academic relevance.



Chapter 1 - Introduction

Problem Identification

The paucity of experts to conduct the BWM questionnaires in the context of standard battles negatively influences the representativity and specificity of the experts' predictions.

Objectives

To find a resolution for the paucity of experts to conduct the BWM questionnaires, this study aims to *understand, test, and examine how the Collective Intelligence of a crowd performs compared to experts.*

Main Research Question

How does the Collective Intelligence perform in comparison to experts when predicting technologies?

Chapter 2 – Literature Review

Theory on factors affecting CI

Diversity & Independence are important features and complement each other to provide a comprehensive view of CI

Theory on CS factors affecting CI

In addition, Crowdsourcing process (i.e. task design/granularity, locating respondents) is discussed because these can also influence CI

Chapter 3 – Conceptual Framework

Hypothesis Development

Utilized the work done in Ch.2 to develop hypothesis regarding relationship between concepts (i.e. Diversity (DIV), Group Size (GS), and Performance (PERF)

Conceptual Framework

Visually representation of the relationship between the three variables (i.e. DIV, GS, PERF)

Chapter 4 – Methodology

Research Approach

Section 1 until section 3 provide a general introduction into the selection of case needed for comparison, survey procedure, targeted audience, boundary conditions and how data was prepared before analysing.

Measurement Level

Assumptions of normality are tested for all the variables

Constructs & Variables

Discusses the constructs of the variables and corresponding attributes.

Conclusion

Selection of appropriate inferential statistics technique

Chapter 5 – Analysis & Results

Diversity, Groups Size, & Performance

Findings One-Way ANOVA & Kruskal-Wallis Test + Multiple Linear Regression (Moderation Effect)

Comparison Performance & Prediction

No inferential statistics applied, only findings based on descriptive statistics

Chapter 6 – Discussion: Significance of the Findings

Key Findings

Answer to the main research question

Implications

Meaning of the results

Limitations

Discussing & evaluating how these limitations influenced the findings

Recommendations

Based on limitations, recommendations for future research are discussed

Chapter 7 – Conclusion: Reflecting this Research

Conclusion

General findings of this research are presented and elaborated

Research Relevance

Discussing the scientific contributions, managerial relevance of this study, and academic reflection

Figure 1.1: Overview thesis structure

Chapter 2

Literature Review

Chapter overview

This chapter discusses the findings of the study of current scholarly literature on CI. This chapter aims to provide a background to the theory on CI.

- *Section 2.1 describes that the idea of CI appears to be true in various circumstances. Still, a group will not always act wisely. Section 2.2. is an extension of the previous section but in the context of crowdsourcing. In other words, the effect the process has on the performance of the crowd.*

This research focuses on investigating an alternative CS solution for the prediction of standard battles. The literature was gathered from two web search engines. Namely, 'Scopus' and 'Web of Science' are known for quality over quantity and information over data. Based on quantifiable data from both search engines, the inclusion and exclusion criteria were formulated. For instance, this data showed when most publications were released and by whom.

The literature review is divided according to the themes found on the concept of CI. In total, twenty themes were identified and evaluated on usefulness and applicability for this research. Selected themes were then divided based on their suitability in the structure of this literature review. This allowed answering the following sub-question:

Sub-question 1: ***Which factors influence the Collective intelligence of a crowd when performing a task?***

Finally, the structure of the literature review is composed of a theoretical and conceptual part. The theoretical framework will define the CI and other concepts and discusses theories and how this research will use these ideas. The conceptual framework will utilize the work done in the theoretical part to develop hypotheses regarding the relationship between the concepts. Resulting in a conceptual model.

2.1 Is the crowd 'mad or wise'?

As the title for this subsection already suggests, leveraging the CI of a group can sometimes lead to 'mad' decisions. Although the idea of CI appears to be true in various circumstances, a group will not always act wisely. One of the solutions would be to enhance the intelligence of an individual. However, this is somewhat challenging in practice, primarily when this person has grown beyond early childhood. In contrast to the individual, it seems highly feasible to do this for groups. Indeed, many researchers are determined to find approaches to explain, enhance or predict the performance of the CI of a group (Suran et al., 2020; Woolley et al., 2015; Yu et al., 2017).

Nevertheless, a comprehensive process theory that explains why some groups are more intelligent than others is still missing (Woolley et al., 2015). However, researchers have been keen on observing CI, and some ideas about the processes of CI groups are recognized. Even without a unifying theory, this study identified and combined several criteria from different researchers. In the literature, there are many other characteristics. However, due to time restrictions and the often ambiguous nature of some articles, this research will only be focussing on two vital features. When these two factors are met, it should result in a crowd acting intelligently. This 'intelligence' results from diversity and independence and is discussed below.

2.1.1 Diversity

According to the Oxford dictionary for languages, diversity is a procedure of engaging with individuals from a broad spectrum of various backgrounds (i.e., social and ethnic), genders, etcetera (Oxford, 2020). Several researchers underlined the importance of this property in respect to the effectiveness of CI (Nguyen et al., 2018; Page, 2007; Salminen, 2015; Surowiecki, 2005; Woolley et al., 2010), arguing that each individual has various perspectives, experiences, knowledge, and interpretations, leading to innovative solutions and better decision-making (Suran et al., 2020). Another benefit for this situation is that mistakes can be canceled out, as pieces of information could counteract or complement each other. Further, several researchers from various disciplines have studied and formulated the criteria for a diverse crowd. Although they originate from various disciplines and applying different perspectives, there are many similarities between them. Thus, several authors and their perspectives on diversity were reviewed and similarities were synthesized.

Surowiecki coined the term CI as the 'Wisdom of Crowds' (WoC). He described that large groups could achieve better results under certain conditions than the most intelligent individual in the group (Surowiecki, 2005). Thus, even if individual estimations are not accurate, the collective estimate can be. According to Surowiecki, five fundamental principles are necessary for the WoC to be successful. The first principle is about diversity. Meaning, each person should have different ways of thinking, amount, and types of knowledge and experiences.

Woolley and his co-authors found evidence for one dominant factor dubbed 'c' that explained roughly 35% of group performance (Woolley et al., 2010). The dominant factor depends on the composition of the group and the interaction of the group members. Thus, their focus was on two influences (i.e., group composition and interaction) on a group's CI. The group composition depends on diversity in members' skills and intelligence. In addition, no correlation was found between c and the intellect of individuals in a group. However, c is positively correlated with the social sensitivity (i.e., recognizing the emotions of others) of group members and the number of females. The latter was possibly enabled because females have a better average social sensitivity than the opposite gender.

In 2010, Krause and his fellow authors termed CI as Swarm Intelligence (SI) (J. Krause et al., 2010). The recognition of SI stems from biology and social insects. In 2011 another paper got published, focusing on identifying opportunities and shortcomings of human SI (S. Krause et al., 2011). Krause and others reported that different factors influence the performance of SI in prediction markets, such as the incentive of contributors for providing information, reliable judgments, minimizing bias, and an appropriate method for managing these opinions (S. Krause et al., 2011). Resulting in three findings: (1) SI benefits vary depending on the kind of issue, (2) performance of the collective or individual can be uncorrelated, (3) adding expertise is not as beneficial as adding diversity to a group.

As for Salminen, his work explores the role of CI in CS innovations. Resulting in several interpretations of CI (Salminen, 2015). Through a systematic review of the literature, various themes from multiple case studies were identified, resulting in a comprehensive framework (Salminen, 2015). This framework consisted of three levels, each with its themes and elements. The second level, 'Macro,' is of interest as it impacts CI, particularly diversity. He refers to diversity in groups of people as variances in learning, cultural upbringings, demographics, and how individuals resolve issues (Salminen, 2015).

Another study focused on activities necessary to promote CI within organizations. By using real-world examples and case studies to develop and propose the required activities to promote CI. This work was primarily influenced by (Page, 2007; Surowiecki, 2005). The authors claim that managers should follow four steps to harness the CI within organizations. The first step involves creating cognitive diversity, described as a mixture of different viewpoints, explanations, heuristics, and projecting techniques (Page, 2007).

Nguyen and his co-authors stated that the collective must abide by four criteria to be intelligent, inspired by Bonabeau's model of Decisions 2.0 (Bonabeau, 2009). The first criteria is diversity in individuals with diverse backgrounds, knowledge bases. The authors claimed this based on an example of weather forecasting. The example implied that even relying on experts for accurate forecasting would be difficult. Nevertheless, allowing multiple individuals to participate in the prediction would lead to additional information and provide various perspectives, which is needed to solve the forecasting problem more efficiently. In addition, diversity entailed "variety in the configuration of participants" and "variety of individual forecasts" (Nguyen et al., 2018). Lastly, they concluded that diversity is an essential criterion for a group to act intelligently based on evidence.

To summarize, diversity entails various components and is an essential condition to leverage the CI of a group of individuals. Due to differences in describing and similarities in meaning, some pieces of information were combined or removed. Thus, for simplicity, in this research, diversity entails involving various individuals with differences in age, (work) experiences, skills, demographics, education, gender, and cultural backgrounds.

2.1.2 Independence

When thinking about independence from a historical perspective, it means to be free of the control of some other person, country, or entity. However, this does not mean that you need to live on your own, but that you have control over your own choices and opinions, based on your private source of experience and knowledge. In essence, this is also what several other researchers have formulated (Lorenz et al., 2011; Matzler et al., 2016; Nguyen et al., 2018; Salminen, 2015). However, it is not the essential criterion for CI to arise, but it is considered the second most important criterion (Lorenz et al., 2011; Page, 2007; Sunstein, 2008; Surowiecki, 2005). There are several reasons why independence is so important. It can help avoid situations such as groupthink (Rosen, 2011), herding, or information cascades that can affect the performance of the CI in a negative way (Lorenz et al., 2011). For instance, with information cascades, users can share information they consider legitimate without the proper proof or understanding, leading to irrational decisions. Next, (Lorenz et al., 2011) showed that minor social interactions could negatively affect the CI performance based on three effects (i.e., social influence effect, range reduction effect, and confidence effect).

In contrast to diversity, (Nguyen et al., 2018) concluded that the impact of independence continues to be contentious because other publications proposed a different view for independence. Namely, if group members are allowed to communicate, their performance increases but decreases when communication is restricted (Matzler et al., 2016; Skaržauskiene et al., 2015; Woolley et al., 2015).

To summarize, diversity and independence are essential features, viewing both features offer a different perspective of CI. In addition, diversity and independence also complement each other to provide a more comprehensive view of the phenomenon. Therefore, measuring these features could help to predict the performance of the collective.

2.2 Crowdsourcing factors affecting Collective Intelligence

In the previous section, two factors that are necessary for a crowd to behave intelligently were discussed. However, the CS process towards leveraging the CI can also influence the performance of the crowd. For the success of a CS system, one of the first activities is the design of a task, which is vital because the initiator explains the task to the crowd through visual presentations and semantics. Therefore, creating an assignment should be made straightforward and unequivocal (Gurari et al., 2016). In addition, (Bhatti et al., 2020) added that visual elements and instructions can influence how workers perform. The following factors were identified and discussed: task design, task granularity, finding the crowd, aggregation.

2.2.1. Task Design

Commencing with task definition, this phenomenon entails describing the task that the initiator offers to the crowd. Accomplishing this task involves components such as information and clarifying the characteristics of the task through keywords (Aris et al., 2016). In addition, (Nakatsu et al., 2014) classified task definition into unstructured and well-structured. For the former there is no defined solution, resulting in ambiguous boundaries of knowledge domains, which will negatively affect the performance of the crowd. In contrast, the latter shows that the solution to the problem had clear boundaries of knowledge domains, which will positively affect the functioning of CI.

2.2.2 Task Granularity

Another factor affecting the successfulness of completing a given task entails the task granularity. Task granularity measures the composability of a task into simple sub-tasks. It is considered an important aspect of task completion, because if a given task is complicated and hard to solve, it would require considerable effort and resources (i.e., cognitive, expertise), increasing the time and cost for the initiator. In addition, the high levels of skills required will also limit the number of respondents willing to participate as potential workers (Bhatti et al., 2020).

On the other hand, if a task is decomposable, it will reduce the risk of failure as the task has become simpler. A task is decomposable into sub-tasks and divided into fine and coarse tasks (Allahbakhsh et al., 2015; Aris, 2017; Bonabeau, 2009). The higher the coarseness of the task, the lower the accuracy and vice versa. Furthermore, other empirical studies showed how errors are affected by task complexity, which makes sense because something straightforward for one person may be hard for others (Suran et al., 2020). However, one can assume that if the complexity of a task increases, this can also increase the opportunity for errors being made. In addition, when a task is very complex, splitting a task into sub-tasks may result in losing context information, hence splitting tasks can be seen as somewhat challenging.

2.2.3 Finding the Crowd

Beyond the previously discussed criteria, finding the crowd members is often accomplished through various platforms such as MTurk and UpWork. Still, finding the crowd members is considered a significant issue in CS systems (Bhatti et al., 2020), especially when the task requirements involve complex and indecomposable tasks. (Schulze et al., 2011) noted that in these situations, randomly involving crowd members can negatively impact the quality of CI. If these specific conditions are required, the task initiator should find individuals who possess this specific knowledge and avoid individuals who do not. Thus, when the task complexity varies, various types of crowd members may be needed to ensure the success of the CS process. Generally, evaluating the crowd participants happens at recruitment through pre-qualification assignments, gold questions, or admission questions (Corney et al., 2010). For example, gold questions (GQ) entail questions where the ground truth is known—allowing to compare the answer of GQ to the actual value to decide the acceptance or rejection of the respondent.

2.2.4 Aggregation

The final factor affecting CI is the validity of the answer and the performance of the overall CS process collected from the crowd performing the given task. Such a mechanism to integrate individual solutions is called aggregation. In general, for an assignment, most experiments apply arithmetic mean or median (Nguyen et al., 2018). For instance, the median is applied to obtain the estimate of a collective group (Galton, 1907). While the arithmetic mean is applied for nearly all observations (Surowiecki, 2005). Previous studies have shown that the geometric mean is superior to those calculated by applying the arithmetic mean and median (Lorenz et al., 2011). As mentioned earlier, CS leverages the CI by outsourcing tasks to a large crowd, leading to multiple solutions for the same tasks while reducing wrong answers and biases (Gao et al., 2015), leading to better innovations and reliable decisions. (Matzler et al., 2016) showed that averaging individual opinions is a good knowledge aggregation technique when predicting markets. Other accumulation methods calculate the results for the individual for each completed assignment (Bhatti et al., 2020). For example, Averaging Output (Avg. O/P) (Surowiecki, 2005).

In the book “The Wisdom of Crowds,” Surowiecki reasoned that a large group could outperform any individual by averaging information. He proposed a relatively simple technique, the Avg. O/P. This technique computes the answers of multiple independent individuals by averaging their output. Resulting in an output close to the actual answer. (Brabham, 2008) confirmed Surowiecki’s claim and showed that aggregating multiple opinions of people could produce an accurate collective answer. This approach is not without any drawbacks as the compensation cost will increase proportionally to the increase of crowd members.

Chapter 3

Conceptual Framework

Chapter overview

This chapter introduces a conceptual framework of the most relevant factors of CI and how these influence the performance of the crowd in a prediction task.

- ➢ *Section 3.1 introduces the motive for the framework. Section 3.2 discusses the process of how influencing antecedents were identified and structured in the framework. Subsequently, Section 3.3 presents the framework showing the relationship between the variables and their influence on performance*.

3.1 Hypothesis Development

CI and CS methods have gained more ground in research, business and other initiatives because an increasing number of people are connected through various devices using the Internet. CS for CI functions on the idea of "Two heads are better than one," where bigger groups perform better at a job than small groups and even experts. The crowd consists of individuals who are geographically dispersed and have various backgrounds. These individuals can make their assessments (in)dependently. The concept of CS allows to reach and collect these individual decisions. When combined with the most suitable aggregation technique, leading to an aggregated answer(s) that is very close to the truth—rendering it a source of intelligence for various purposes, such as predictions.

As mentioned previously, some use CS and CI interchangeably, others proposed several definitions but with subtle differences. However, combining both concepts is required to solve a task by outsourcing and utilizing distributed human capabilities online. Many researchers attempted to explain the phenomenon of CI without a unifying theory explaining the underlying mechanisms. Thus, because of these various but similar definitions and underlying mechanisms, CS and CI are challenging to understand and explain.

When looking into the dimensions that can influence whether the aggregated outcome is close to the truth, some researchers combine the process and the crowd, resulting in a list of factors. On the other hand, other researchers focussed only on factors impacting CI or CS. In both cases, many similarities were present. This research however will focus on three variables: 'diversity' (DIV), 'group size' (GS), and the performance (PERF) of the crowd. Specifically, this research investigates the differences in PERF of a collective when predicting the outcome of a standards battle compared to that of experts.

In the next section, the variables DIV, GS, and PERF are elaborated, from which several hypotheses are derived. Subsequently, a conceptual model is presented, showing the relationship between the three variables.

3.1.1 Diversity (DIV)

Several researchers have stressed and explained the importance of DIV and its effects on the PERF of a crowd to act intelligently. (Woolley et al., 2010) examined three factors (i.e., social sensitivity, equal conversational turn-taking, proportion females) and found these correlated significantly with CI. However, the former and second factors of the three were statistically insignificant after controlling for social sensitivity. Others focused on establishing relationships between DIV and PERF and between cognitive diversity and PERF. Often DIV is described as a group with various individuals with differences in age, (work) experiences, skills, demographics, education, gender, and cultural backgrounds. While (Nguyen et al., 2018) categorized DIV as the 'variety in the configuration of participants' and 'variety of individual forecast.' As (Page, 2007) described, cognitive diversity consists of diverse perspectives, interpretation, heuristics, and predictive methods. Reynold and David Lewis researched the relationship of (cognitive) DIV and PERF worldwide for more than 12 years. They found strong support for the correlation between cognitive diversity and PERF, but no correlation between DIV and PERF (Reynolds et al., 2017).

Based on previous statements, it can be argued that DIV of the crowd is not the magic bullet that leads to more accurate predictions. Nevertheless, there are several reasons why DIV is of a more significant benefit for this research than cognitive DIV. First, unlike ethnic or gender variety, DIV based on the cognitivist view is hard to detect and measure (Reynolds et al., 2017). Second, improved DIV through various socio-economic, cultural, and educational backgrounds is a helpful way to improved cognitive DIV, especially in the case where respondents previously came from a limited range of backgrounds as with experts. Third, DIV allows access to the broadest pool of talents available.

As Surowiecki claimed, individual predictions may not be accurate. However, a collective prediction can be because a set of diversified actors has many individuals with various pieces of knowledge, perspectives, interpretations, and experiences, resulting in information that would either counteract each other's mistakes or complement the correct answer. Through this process, mistakes cancel each other while complementing the correct answer. Therefore, the more diverse the crowd members are, the more information is canceled out and matched, which results in a aggregated answers that should be closer to the truth than with less diversity. In conclusion, this deduction leads to the following hypothesis:

H 1. *Diversity (DIV)* ⇧ ➔ *Performance (PERF)* ⇧
More diverse groups of individuals will have a strong and positive effect on the performance of the crowd

3.1.2 Group Size (GS)

Just as the saying goes, "many hands make light work." In this case, more crowd members means more information, leading to better performance. Hence, leading to better decisions that will make the work lighter. Researchers have several contradicting differences in what affects the number of crowd members and the Optimal Group Size (OGS). In other words, the quantification of the phenomenon lacks conceptual clarity, as will be shown.

Some scholars claim that the effect of GS is mainly dependent on the task that the initiator proposes. Particularly the task complexity as laid out by (Bhatti et al., 2020). (Schulze et al., 2011) noted that in a complex task, randomly involving crowd members will negatively impact the PERF of CI. In addition, the study of SI by (S. Krause et al., 2011) tested three samples with varying difficulties. Findings indicated that the highest PERF score was with a medium-sized problem with a swarm size of 25 agents.

In contrast to these studies, (Carvalho et al., 2016) empirically investigated the effect of task complexity and incentives concerning the OGS. They found support that the optimal number is independent of the underlying task and motivation. They claimed that the OGS for each task is around 10 to 11.

Lorenz approached the CI from a different angle. He explained the phenomenon based on the definition of an ISO standard resulting in two measures (Lorenz, 2021). The second measure, 'fraction of outperformed estimates,' quantifies the OGS based on the expectation of systematic biases in the estimation tasks. In other words, the OGS appears when estimates from a crowd show high trueness (low collective error) and low precision (high variance). As for the worst situation, the initiator observes a false consensus characterized by low trueness and high precision (Lorenz et al., 2011).

Other scholars also occupied themselves with the question of what the OGS would be. One would assume that when more individuals are added, the PERF will become better. In practice, evidence contradicted this assumption, indicating an upper limit after which the PERF does not increase, and in some cases even decreases. The study of (Hashmi, 2005) explained this by showing a curvilinear relationship between the GS and their PERF. Hence, confirming there is an upper limit. In addition, larger groups are outperforming groups that are smaller independent of the task. Still, the slope decreases when GS increases (Green, 2015). This study further concluded that a group of 10 performs better than any individual and that groups of 20 perform marginally better than those with 10 individuals, therefore confirming the view of (Carvalho et al., 2016). To conclude, the deduction leads to the following hypothesis:

H 2. ***If the crowd size increases until the Optimal Size, the crowds' Performance also increases. After this point, the relationship is negative.***

3.1.3 Moderation Effect

As shown previously, two key characteristics of the crowd are GS and DIV. However, both these characteristics do not impact the crowd performance independently. Instead, it is argued that the impact of GS and DIV on the PERF of the crowd must be taken into account together (Oliver et al., 1988; L. Robert et al., 2015). To put it another way, crowd DIV depends on GS and vice versa (L. P. Robert et al., 2017). When the crowd members are more diverse, an increase in GS will lead to an improved PERF. However, when the crowd members are less diverse, an increase in GS should deteriorate the PERF of the crowd. Arguments related to why this is the case are explained in the following section.

In the section above, it was shown that "many hands make light work" because an increase in GS means increased information, resulting in an improved PERF. According to (Page, 2007), this is only valid if the additional crowd members bring supplementary knowledge, perspective, interpretations, and unique experiences. This explanation resonates with the first key characteristic, DIV, because it reveals these polarities in skills and knowledge (Chen et al., 2010). Thus, if DIV in a crowd is low, an increase in GS should negatively affect the PERF of the crowd. This makes sense because you will be adding individuals with the same DIV, leading to herd-like behavior and its negative consequence. On the other hand, if DIV in a crowd is high, an increase in GS should positively affect the PERF of the crowd because mistakes get canceled while reinforcing the correct answer.

Amongst researchers several contradicting differences were noticed if the OGS exists. If so, differences in quantity were also present. This research assumes that the OGS indeed exists, based on one claim (Page, 2007). In addition, to strengthen this claim, an example is shown based on a constructed thought experiment. Imagine a situation where it is possible to contact and persuade every human in the world to participate in a survey for a specific task. In this case, it is impossible to add more DIV by adding more people because there is nobody to add. However, one option would be possible by re-educating people or by providing them with new experiences, etcetera. However, this option would require a significant amount of effort and would not be feasible in reality. Nevertheless, let's say that re-educating people was successful, leading to a more diversified crowd. Even in this case, there should be an upper limit for the simple reason that every human may indeed be capable of acquiring and grasping all the different experiences in a single lifetime. If assumed this is the case for everyone, in the end, this will result in a homogenous group of people, and thus, diversity diminishes.

Based on the contradicting claims about the number of participants needed to leverage CI of a crowd and the ambiguous nature of the upper limit. The following is proposed:

H 3. If Diversity (DIV) is high ↗ Group Size (GS) ⇧ ➔ Performance (PERF) ⇧
Crowd Diversity moderates the relationship between Crowd Size and Performance. Size increases are positively related to performance when crowds are more diverse than with more similar crowds.

3.2 Conceptual Model

As previously shown, arguments and views were combined to develop three hypotheses, resulting in a conceptual model shown in the following abstract model.

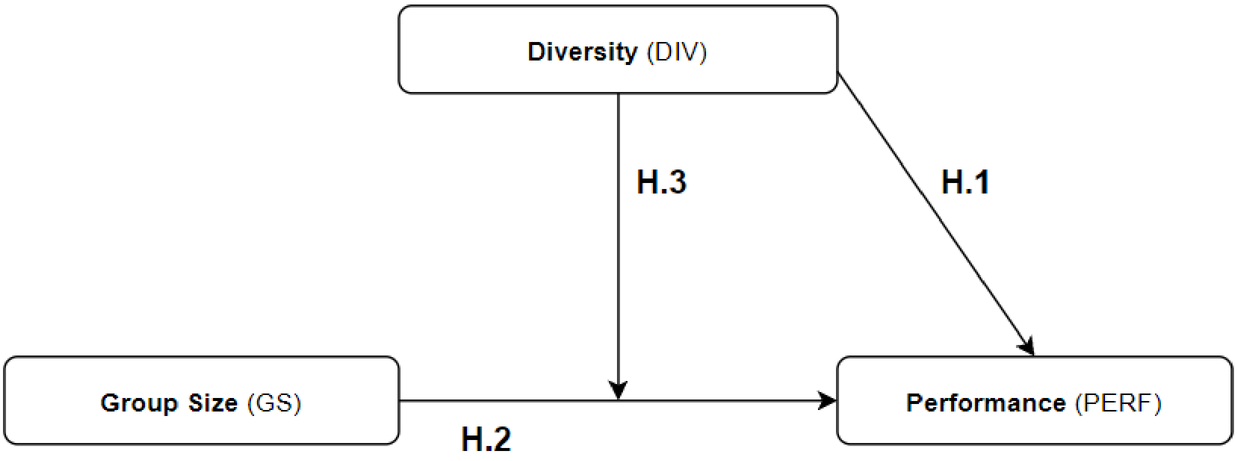


Figure 3.1: Research Model – Group Size (GS) & Diversity (DIV) Relationship with Crowd Performance (PERF)

Chapter 4

Methodology

Chapter overview

This chapter explains this research's organization and execution, allowing fellow academics and other readers to evaluate this study.

- *Section 4.1 is a general introduction to problem identification, results from literature review, and the case of GB VS DD to exam the proposed hypotheses from the previous chapter. Section 4.2 explains the survey procedure, targeted audience, and letter of informed consent. Section 4.3 entails how data was collected and managed in this study by explaining the boundary conditions and preparation of the data before analyzing. Subsequently, Section 4.4 discusses the constructs of the variables and their attributes. Section 4.5 elaborates how the raw data was prepared and represents some descriptive statistics about demographics. Section 4.6 present the assumptions of normality for all the variables. Finally, section 4.7 provides the overall conclusion drawn from the previous sections.*

4.1 Intro

This section explains how the research got executed to allow readers to evaluate the reliability and validity of this study. The strategy and methods are primarily dependent on the sub-questions and will therefore roughly follow the same structure to improve the understanding and readability.

4.1.1 Problem identification & Research Design

Like every scientific research, the first step is to identify the problem through a preliminary literature review. After this, a research design was formulated based on the research objectives and questions.

4.1.2 Literature Review

A literature review has been conducted, providing the required theoretical background to help answer our research questions. Based on the literature review, three deliverables were created: Exploring the under-researched topic of CI in the context of MCDM problems, systematically describing the characteristics affecting CI, and establishing a relationship between DIV, GS, and PERF.

4.1.4 Case Description

As stated before, to establish an 'ante' or 'post' battle prediction, a combination of the work of Geerten and his framework of 29 factors and the BWM developed by Jafar is applied (Rezaei, 2015a; v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, 2011). The process of such a prediction relies on experts' opinions, and in some cases on secondary data. This process entails three steps: selecting relevant factors, weighing the relevant factors, and determining a performance score for every technological solution proposed. This research focuses only on scoring the relevant factors of the alternative solutions using the crowd and leveraging their CI.

Next, the claim that CI can perform similar to experts and initial hypotheses were tested. One prior research on standard battles was selected—namely, the study involving the battle of WTT. More specifically, the GB versus DD technology (van de Kaa et al., 2020). Because of several factors, such as the outcome of the battle was predicted by experts but where the ground truth is unknown. Doing so allowed validating if the crowd performs differently than experts in prediction tasks. Another aspect that influenced this choice was the availability of data of the prior study, but also time and monetary expenses as these resources were limited.

4.2 Method

This section elaborates on the justification of the research method in this project to allow other researchers to replicate this study and its conditions and verify if this research is valid. This study is considered a quantitative exploratory research project following deductive reasoning from the literature review. Concluding from the literature review, interest in the effects of DIV and GS on the PERF of the crowd is deemed relevant. However, this relationship was not investigated before in standard battles and solving the MCDM problem by applying the BWM. An observational study with a quantitative method is required to examine and test the initial hypothesis that DIV and GS affect the PERF of a crowd. Doing so allowed to examine the relationship between GS, DIV, and PERF by combining secondary data (i.e., prior research on standard battles) and primary data. The latter entails descriptive data obtained through observations without the intervention of the researcher. The gathering of data was possible because the traditional BWM got converted into an online distributed survey.

4.2.1 Survey Procedure

In this research, an online survey strategy was applied. This approach studies the sampling of individual units from a population through a questionnaire of one or more questions that respondents answer.

Furthermore, a cross-sectional online survey was administered, consisting of several electronic questionnaires. The survey was hosted at Qualtrics, an online survey design service, allowing to design, prepare, and administer the questionnaires (Qualtrics, 2021). The respondents or crowd members completed the survey in their private time with the help of an Internet device, hence the questionnaire was self-administered.

4.2.2 Target Audience

This study targeted the entire population dispersed around the world that had access to the Internet through a device (i.e., smartphone, computer, laptop, tablet, etcetera). At the beginning of 2021, there were 4.66 billion Internet consumers active worldwide (Almomani et al., 2021), representing 59.5 percent of the global inhabitants. In addition, of this total, 92.6 percent (4.33 billion) gained access to the Internet through their mobile devices (Almomani et al., 2021). This population frame is substantial due to the international nature of the theory on CI.

4.2.3 Sampling Size

Appropriate sample size is larger than the minimum size of 100 required because of the limited time and capital. However, these limiting factors usually advises choosing a number closer to the minimum size. This was chosen because it was expected that the respondents would provide various answers, meaning that the results need to be as accurate as possible. Making it is necessary to move towards a number that lies closer to the maximum size. In addition, a bigger sample size was required because the sample gets divided into many groups.

To conclude, this study is interested in carrying out an observational study using simple random sampling from a population of over 4.66 billion Internet users. Were the acceptable margin of error for the estimate should be within five percent of its actual value. Since the population is finite yet large (more than 5000), a sample size of more than 100 respondents will allow the study to determine the association between diversity, size, and performance with a plus and minus five percent confidence interval.

4.2.3 Letter of Informed Consent

To ensure that the participants were interested in taking part in this study, a letter of informed consent provided them with additional information on the project, ensuring the ethical gathering of the data. This message got presented at the beginning of the questionnaire.

> *"Welcome to the questionnaire on the impact of 'Collective Intelligence' of a crowd and its performance in comparison to the performance of experts. This project examines the relationship between crowd diversity and the size of the crowd and its outcome. More specifically, to see if a random crowd can determine a similar technological prediction compared to experts. You will be presented with information and questioned about your views about a particular technology by grading various factors that influence this technology. This score is then averaged, which is then used to rank the technologies. Based on this ranking, a prediction is then made on which technology will dominate the market. Please be assured that your responses will be kept confidential and anonymous."*

After this paragraph, the researcher got introduced to a short explanation of standard battles, a list of categories and factors used along with their definitions, and the used measurement scale. The next page showed how the crowd members were informed about collecting and handling data during the project and completion.

> *"This questionnaire is part of the master thesis project by Igor Djordjevski, an MSc student from the Delft University of Technology, the Netherlands. During the project, the data will be stored on a ProjectDrive at the Delft University of Technology. The data will be published in Igor Djordjevski master's thesis to assess the quality of the research. After completing this project, the gathered data, analysis, and results will be publicly available in the Delft University of Technology."*

The participants were told about the duration of the questionnaire, how to contact the researcher and provide consent. Which got presented as follows:

> *"This questionnaire is part of the master thesis project at the Technical University of Delft by Igor Djordjevski.*
>
> *The survey is about grading two current wind turbine technologies (i.e., Gearbox & DirectDrive) and consists of 42 questions. The first couple of questions are about demographics and are multiple-choice. The remaining question needs to be answered based on a 9-point Likert scale. The questionnaire should take around 4-8 minutes to complete. Your participation in this project is considered voluntary. Please be assured that your responses will be kept confidential and anonymous.*
>
> *You have the right to withdraw during the study for any reason and without any prejudice. If you would like to contact the researcher of this project, please send your e-mail to the following address: i.djordjevski@student.tudelft.nl*
>
> *In advance, I thank you kindly for your participation in this research study."*

If someone decides to select the option of consent, they are allowed to fill in the questionnaire. Otherwise, appreciation for their time and effort was shown.

4.3 Data Management & Collection

The survey was distributed through one channel dubbed MTurk. In Appendix A, a complete overview of the survey is depicted. MTurk is a crowdsourcing marketplace that makes it easier to outsource tasks to a distributed workforce who can virtually perform them (Amazon, 2021; Gyulavári, 2020). However, this service does require financial compensation. The compensation is roughly €1,22 per completed questionnaire. Table 4.1 shows the results.

Table 4.1: Platform distribution

PLATFORM	FREQUENCY (N)	PERCENTAGE
MTURK	200	100%
TOTAL	200	100%

4.3.1 Boundary Conditions

Before analyzing, the gathered data was prepared to satisfy the boundary conditions, as shown in Table 4.2. As for the 'GQ,' this entails questions where the ground truth is known—allowing to estimate the validity of the given answers without analysis by comparing the answers of the crowd members to the truth allowed to include or exclude some participating individuals.

Table 4.2: Boundary conditions

BOUNDARY CONDITION	DEFINITION
Consent	*Crowd members have provided their consent to the questionnaire in the pre-screening question*
Completion	*Crowd members have a 100% completion rate of the questionnaire*
No missing values	*Crowd members have no missing values in all the questions*
Gold questions (GQ)	*Crowd members who have answered these questions correctly*

To conclude, a summary is provided in the table below, illustrating the total removal of crowd members that did not meet the boundary conditions. This research excluded a total of 63 individuals.

Table 4.3: Summary exclusion

BOUNDARY CONDITION	REMOVED
Consent	0
Completion	0
No missing values	0
Gold questions (GQ)	63
Total	**63**

4.3.2 Respondents

After collecting the data, respondents who did not satisfy the boundary requirements were excluded—resulting in a sample of 137 respondents for analysis.

4.4 Constructs & Variables

In quantitative research, it is necessary to bring the findings from the literature review, constructs, and variables together. In general, constructs are seen as a mental abstraction, helping to explain how and why certain phenomena behave the way they do. However, constructs can vary significantly in their complexity and are seldom directly observable. Hence, conceptual clarity is required to establish research of high quality. The second step entails translating the abstract construct to a concrete variable that can be measured and tested. Because the variables also have their attributes (e.g., variable = gender, attributes = male/female) that can be measured. The aim of operationalizing a definition is to describe how to measure the characteristics of a construct precisely.

The section below elaborates the two constructs (i.e., Collective Intelligence and Performance) and their variables and attributes. However, that section will not explain why and how it was operationalized because Chapter 3 elaborated this in-depth.

4.4.1 Collective Intelligence (CI)

This construct was viewed from several different perspectives, which followed many definitions and similar concepts. In other words, the construct CI is more complex, and thus, more challenging to understand and measure. Nevertheless, the following integrated definition is accepted, as it reflects various viewpoints of the creator, group members, and process in solving problems:

> ***"Crowdsourcing is an online distributed problem-solving paradigm, in which an individual, company, or organization publishes defined task(s) to the dynamic crowd through a flexible open call to leverage human intelligence, knowledge, skill, work, and experience." (Bhatti et al., 2020)***

VARIABLES & ATTRIBUTES

Even without a unifying theory on why some groups are more intelligent than others, this study identified and combined several criteria from different researchers, due to time restrictions and the often ambiguous nature of some articles, thus this research focusses only on 'DIV' and 'GS.' The control variable, in this case, is independence. Meaning that the individual members of a contrived group independently complete the given task based on their own knowledge and experience. This was controlled because it was assumed that the only way to verify this variable was through a one-sided trust relationship between respondents and the researcher. Hence, the variable was designated as a control variable because there is no proper way to verify this condition.

Several researchers from various disciplines have studied and formulated the criteria for a diverse crowd. Although they originate from various disciplines and apply different perspectives, there are many similarities between them. This research therefore reviewed and synthesized several authors on their perspectives on DIV. Consequently, the representation of attributes of the DIV variable involves various individuals that have differences in gender, age, degree, job, and nationality. Table 4.4 indicates how these attributes were measured.

Table 4.4: Measurable attributes of the diversity (DIV) variable with corresponding labels

	GENDER	AGE	DEGREE	JOB	NATIONALITY
RANGE	1-3	1-8	1-7	1-8	1-7
LABEL	Male	<20	>GED	Working	Afrika
	Female	20-24	GED	Self Employed	Antarctica
	Other	25-29	No Degree	Laid-off	Azia
		30-34	Associate Degree	Looking for work	Australia / Oceania
		35-39	BSc	Retired	Europa
		40-49	MSc	Disabled	North America
		50-59	PhD	Other	South America
		60<		Prefer not to say	

In addition, the table above was combined with the Simpson's index (SIMPSON, 1949). This index of DIV, also called the phylogenetic index, is a numerical expression indicating the amount and distribution of different species. The computation of this index is as follows:

$$D = \sum_{i=1}^{R} \left[\frac{n_i(n_i-1)}{N(N-1)}\right] \tag{1}$$

Where 'ni' represents the number of people in species 'i,' and 'N' is the total sum of species. The Simpson's D can take any value between zero and one. Where zero is representative for infinite DIV, and one represents the scarcity of DIV. In other words, if the value of D increases, DIV decreases.

For example, to compute this index for a theoretical case with three species.

Species	No. of Individuals (Group A)	No. of Individuals (Group B)
Male	6	4
Female	3	4
Others	3	4

First, calculate N, thus, NA = 6 + 3 + 3 = 12 and NB = 4 + 4 + 4 = 12. Now that the amount of individuals for each species is known, the Simpon's index can be computed as follows:

$$D_A = \sum_{i=1}^{R} \left[\frac{n_i(n_i-1)}{N(N-1)}\right] = \left(\frac{6(5)}{12(11)} + \frac{3(2)}{12(11)} + \frac{3(2)}{12(11)}\right) = \frac{42}{132} = 0{,}318$$

$$D_B = \sum_{i=1}^{R} \left[\frac{n_i(n_i-1)}{N(N-1)}\right] = \left(\frac{4(3)}{12(11)} + \frac{4(3)}{12(11)} + \frac{4(3)}{12(11)}\right) = \frac{36}{132} = 0{,}273$$

DB is considered more diverse than DA because the group members are more consistently distributed amongst the three species.

In contrast to DIV, the size of the crowd is a variable that is easy to understand and measure. Therefore, the justification is relatively straightforward and does not require any further explanation because of the assumption that others would choose the same variable and attribute. In this study, the variable of GS has only one attribute. Namely, the number of people in a group that are participating in the survey.

4.4.2 Relative Performance (RP)

To get a clear understanding of the complexity of PERF, the approach as depicted by (Wagner et al., 2010) is discussed. Here, the measurement of the PERF of the collective intelligence entails the question of "*how close to the true value is close enough*?". Answering this question is somewhat challenging because if the PERF of the collective is 40% of the prediction made by experts (i.e., actual value), the PERF is regarded as poor. However, if the PERF is 60% of the actual value, then the CI of the crowd is considered superior. Subsequently, describing the answer to the previous question cannot be described in absolute terms.

Thus, this research applies a CI metric as described in the study of Wagner based on statistical reasoning (Wagner et al., 2010). Namely, by calculating the distance between the 'true value' and the outcome of the collective. Doing so captures the relative PERF of the crowd. In addition, based on the t-statistics, the collective outcome is considered expert-like if the confidence interval (proximity to the actual value) would be around the sample mean, allowing no more than p = 0.05 likelihood for this to happen. To conclude, the PERF is defined as the differences between the solution made by experts and the crowd. Hence, the rest of this paper will be about relative performance (RP).

4.5 Measuring Variables

Until now, the overall strategy for the data collection and management was shown and discussed. In this section, it shows how the collected data was processed to be valuable and useful. Secondly, descriptive statistics for the demographics is displayed. The next step is to determine which statistical test suits the variables GS, DIV, and PERF. This was done by identifying the distribution of the variables.

4.5.1 Preparing Data

As stated earlier, 200 responses were collected, but only 137 remained after pre-qualifications. The data itself was collected through Qualtrics and converted into an Excel and SPSS file. The former was required due to calculations that needed to be done manually.

It was necessary to check if the respondents correctly graded the criteria after selecting the best and worst criteria. For instance, when comparing the best against the best criteria or the worst against the worst criterion, one would assume that both will get an identical score because there are no differences. Likewise, the score will indicate the most significant difference when comparing the best against the worst criteria. This inspection was required because Qualtrics did not provide a function to automate this basic logic. The color red was given to the worst criteria, and green was the best factor to circumvent this issue. After a manual inspection of every respondent for every alternative technology, most people did not get this right. The consequence of this would result in a consistency ratio that was much higher than the threshold value. Manually altering this basic logic led to a consistency ratio below the threshold value.

The (Solver Linear BWM, 2016) Jafar Rezaei was applied to calculate the consistency ratio and PERF score. The latter was then multiplied with the weights per criteria that the experts established. Resulting in an overall score per alternative technology. However, to obtain the RP of the crowd, the grade of the experts was deduced from the overall score. The variable RP indicates the distance in PERF of the crowd compared to that of experts.

In this research, 'groups' were contrived from the sample that completed the prediction task. The group members employ their expertise to carry out the given task. In other words, all group members perform the same activity. In the end, their results are pooled randomly with varying group sizes.

The next step involved simplified random sampling, which was done twice for every independent variable (i.e., GS, DIV). For the former, it is required to investigate if the number of people in a group affects their RP. Hence, all the 137 respondents were randomly assigned once to one of the six groups. If one group was complete, the following respondent would be assigned into the next group, etcetera. The size of groups 1, 2, 3, 4, 5, and 6 was respectively 5, 10, 15, 20, 30, 40, resulting in 120 cases. However, the division of the respondents amongst six groups would not provide enough data based on their DIV scores, leading to a total of six scores. Therefore, simplified random sampling was done again for every Nth group. In other words, the whole sample was used multiple times across these groups, resulting in multiple subgroups, each with a DIV score. Resulting in groups of size 5, 10, 20, 30, and 40, with 26, 13, 6, 4, and 3 subgroups for comparison. This comparison was not made between the GS because respondents are similar for every group, thus leading to no differences. Finally, the DIV score for every group was manually calculated.

4.5.2 Descriptive Statistics

This section summarizes the characteristics of the data set by showing the frequency distribution of the respondents' demographics, which helps in understanding the collective properties of the elements in the sample.

As stated before, the data gathered about demographics entailed gender, age group, educational background, employment, and geographics, as shown in Figure 4.1.

To summarize the findings, the order as is depicted in the previous sentence is followed. The respondents selected only two out of five options for gender. These were male and female. The proportions are 101 (73.7 %) and 36 (26.3 %) respondents respectively. Age consisted of nine groups ranging from below 18 until 85 or above. The sample represented only six groups, where the gross majority was between the age of 25 – 34 (40.2%) and 35 – 44 (36.5%). As for employment, there were eight options which were all selected by the sample. However, ‘employment full time’ was selected 118 times, meaning 86.1% of the respondents chose this answer. When looking into the spread of respondents based on their geographics, only four out of seven options were selected. The majority resided in North America at 54%, after which was Azia with 21.2% and third with South America at 19.7%. In total, 73.7% of the respondents originated from the Americas. Finally, the educational background had seven categories, all of which were selected. Nevertheless, the proportions were not equally distributed. Namely, 77 (56.2%) people had a bachelor as their education. Associate degrees and masters were close together, but the former was second with 19% and the latter 15.3%. In other words, 90.5% of the respondents came from a higher educational background.

To conclude, based on demographics, the core characteristics of the sample can be described as male, relatively young, highly educated, full-time job, and lives in the Americas. In hindsight, this makes sense because the topic involves technology which generally attracts more males than females. Similar sensemaking for geography, age, and education was used. Namely, the questionnaire was put online at MTurk, which is part of the company Amazon. Most of their customers reside in The Americas, which probably resulted in a higher proportion of respondents from this area. In addition, some older adults may skip this type of questionnaire due to their online nature or simply because they have no interest or knowledge of its existence. Resulting in the core characteristics found in the sample.

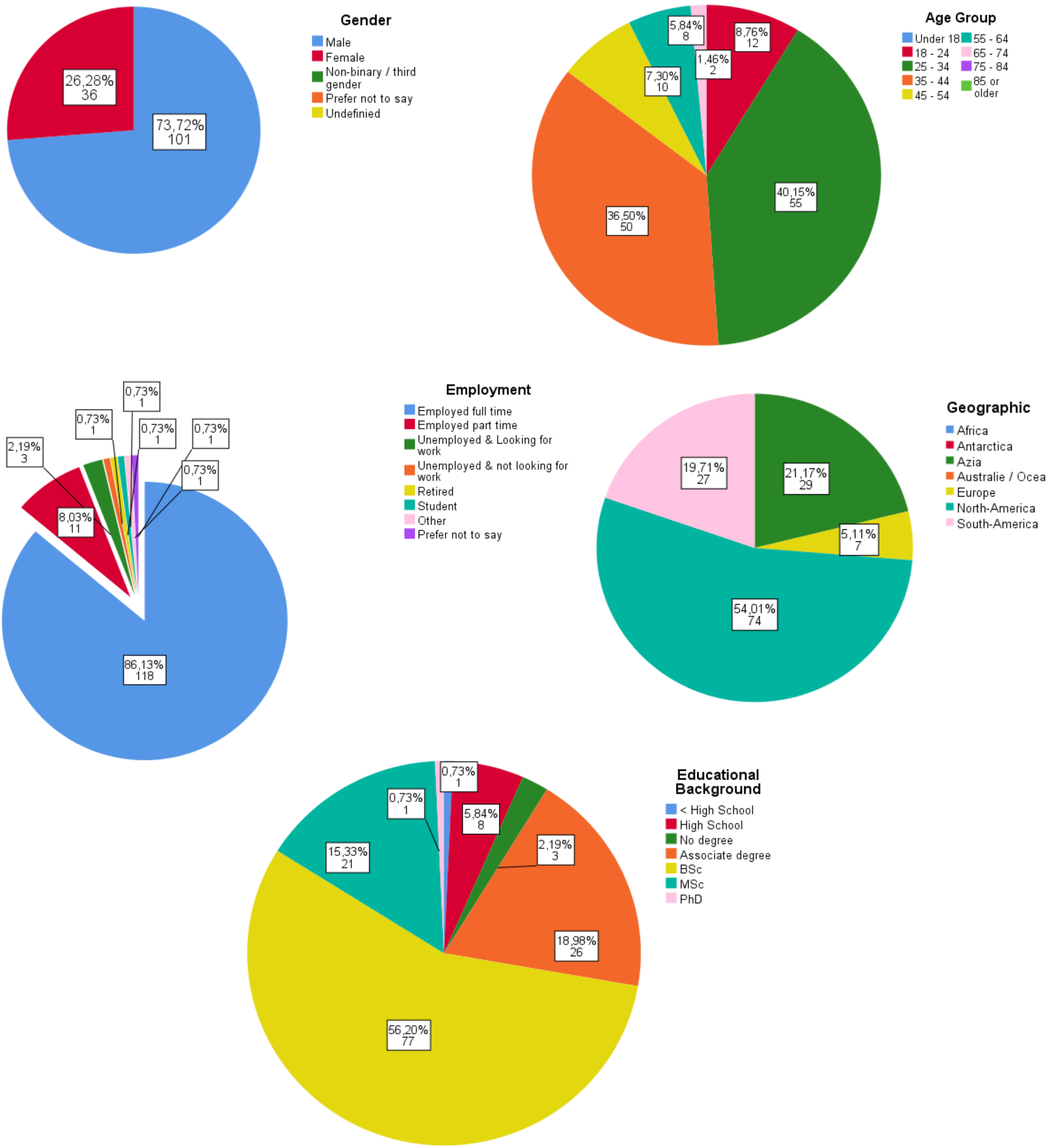


Figure 4.1: Pie chart showing distributions of respondents demographics

4.6 Measurement Level

In the previous parts, the preparation of the data and sample's characteristics were discussed. In this section, the distribution of the independent variable (IV) (i.e., GS & DIV) and dependent variables (DV) (i.e., RP for GB & DD) are identified. In other words, to decide if to proceed with non- or parametric tests, the variables are tested if they satisfy the normality assumptions required for parametric testing.

For the IV, the skewness and kurtosis values and their standard error are investigated. In essence, Skewness measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution. Thus, the Z-value is obtained by multiplying the value of skewness and their standard error (likewise for kurtosis). These values combined indicate the shape of the data. If these values are between -1.96 and 1.96, it can be concluded that the data is in the range of a normal distribution (Cramer, 2003; Cramer et al., 2004; Doane et al., 2011). If exceeded, the outliers were detected by applying the Mahalanobis distance (Md). The latter technique indicates which respondent shows a significant Md when compared to others. Consequently, more influence is exerted on the slope when the leverage is high. Subsequently, the outliers were replaced accordingly.

As there are no perfect tests for all forms of non-normality, the Shapiro-Wilk test (SWT) was also utilized. The Kolmogorov-Smirnov test (KST) was not, because it is less powerful compared to the SWT. This means that to a lesser extent, discards the premise that the data is normally distributed (Mohd Razali et al., 2011). When there is no normal distribution, and the significance estimate is above p = 0.05, the null hypothesis is accepted and vice versa (SHAPIRO et al., 1965).

For the DV, the Levene's test was applied to assess the equivalence of variances for a variable computed for two or more sample sets. It analyses the notion if the sample variance is identical, also called homogeneity of variance. For instance, if the p-value of the test is less than some significance level (i.e., p = 0.05), indicating the disparity between the variances. Therefore, rejecting the null hypothesis and applying more generalized tests free from homoscedasticity assumptions (i.e., non-parametric test).

4.6.1 Assumptions – Group Size & Relative Performance

For this section, no graphs or tables are presented. However, Appendix D provides a complete overview of this part.

For the IV GS, the SWT for F = 0.871 (df = 5) found Sig. of 0.001 (p << .05), which indicates that the IV does not satisfy the normality assumption that is required to perform parametric tests. Furthermore, the skewness value of -0.290 (SE = 0.221) indicates a Z-value within the threshold range. Thus the IV does follow a normal distribution. However, the kurtosis value of -1.235 (SE = 0.438) shows a Z-value below -1.96. Hence, the IV does not follow the normal distribution.

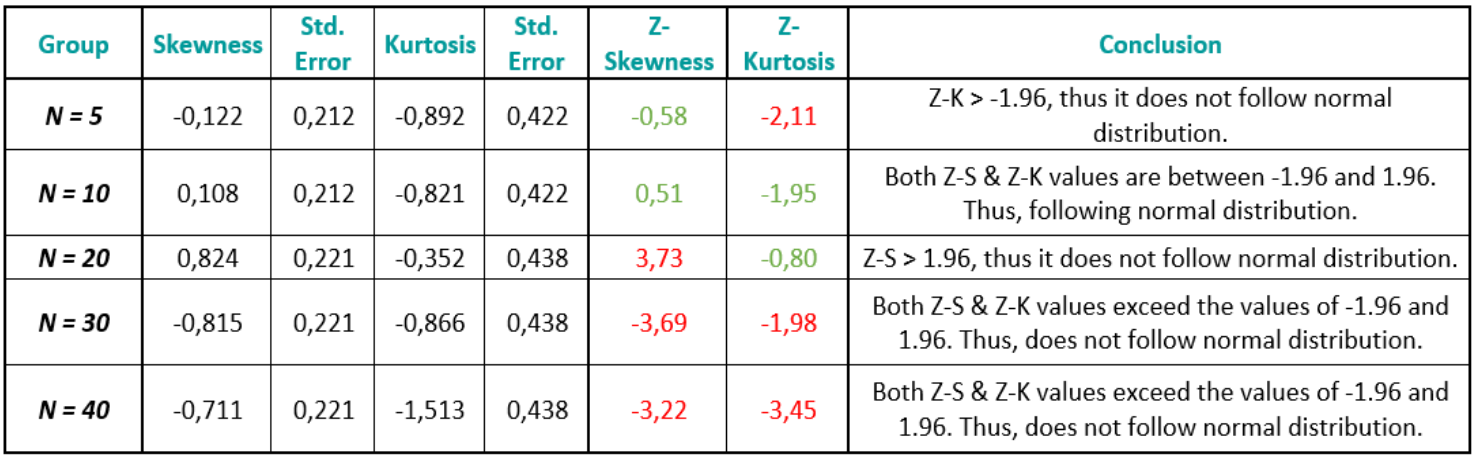
For the DV, two variables required testing. Namely, the RP for the GB and the DD. The Levene's test showed for the GB a Levene's statistic (Ls) of 1.237 (df2 = 114) with a Sig. of 0.296 based on its mean. As for the DD, the Ls of 1.740 (df2 = 114) with a Sig. of 0.131 based on its mean. In both cases, the test found no significant values for the RP, which means that the DV shows indications for homogeneity of variance.

4.6.2 Assumptions – Crowd Diversity & Relative Performance

In this section, DIV and RP for the GB and DD for every Nth group is tested. Table 4.5 shows the skewness and kurtosis values and their conclusion. It was concluded that the IV follows a normal distribution, but only for the group of 10. The other groups did not have a normal distribution.

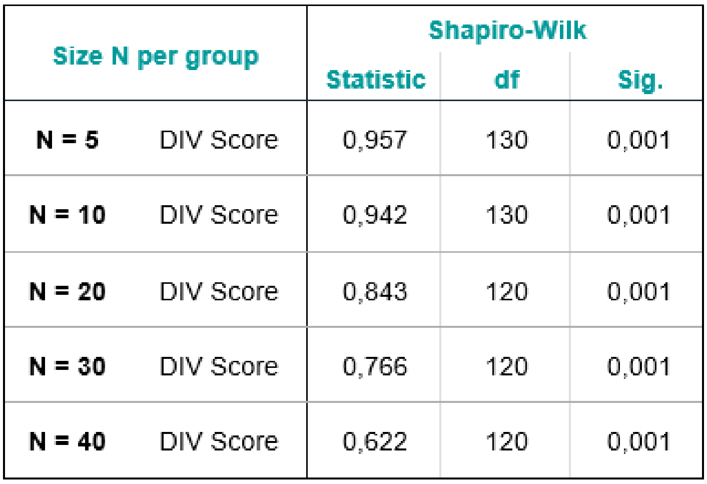
Table 4.6 shows the results for the SWT for all the group sizes. These results indicate a significant value of p = 0.001 for all the groups, which is far below the threshold value of p = 0.05. Thus, this test indicates that all the groups based on their DIV score do not satisfy the normality assumption, which is quite similar to the conclusion of the previous test. The absolute values and tables can be found in Appendix C for both these tests.

Table 4.5: Skewness & Kurtosis for diversity (DIV) score



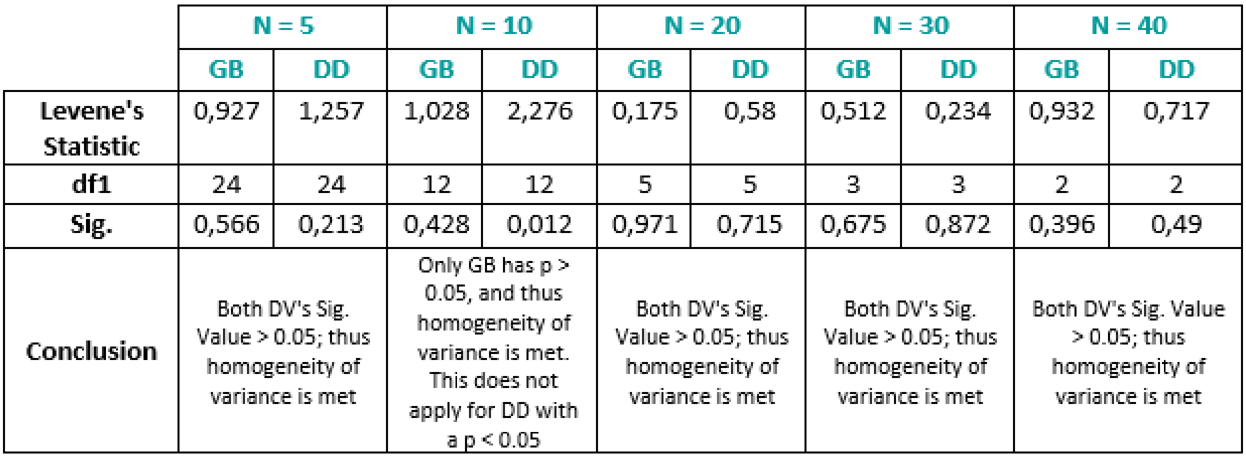
Group	Skewness	Std. Error	Kurtosis	Std. Error	Z-Skewness	Z-Kurtosis	Conclusion
N = 5	-0,122	0,212	-0,892	0,422	-0,58	-2,11	Z-K > -1.96, thus it does not follow normal distribution.
N = 10	0,108	0,212	-0,821	0,422	0,51	-1,95	Both Z-S & Z-K values are between -1.96 and 1.96. Thus, following normal distribution.
N = 20	0,824	0,221	-0,352	0,438	3,73	-0,80	Z-S > 1.96, thus it does not follow normal distribution.
N = 30	-0,815	0,221	-0,866	0,438	-3,69	-1,98	Both Z-S & Z-K values exceed the values of -1.96 and 1.96. Thus, does not follow normal distribution.
N = 40	-0,711	0,221	-1,513	0,438	-3,22	-3,45	Both Z-S & Z-K values exceed the values of -1.96 and 1.96. Thus, does not follow normal distribution.

Table 4.6: Shapiro-Wilk Test (SWT) for diversity (DIV) score for every Nth group



Size N per group		Shapiro-Wilk		
		Statistic	df	Sig.
N = 5	DIV Score	0,957	130	0,001
N = 10	DIV Score	0,942	130	0,001
N = 20	DIV Score	0,843	120	0,001
N = 30	DIV Score	0,766	120	0,001
N = 40	DIV Score	0,622	120	0,001

Table 4.7: Levene's test for relative performance (RP) score of Gearbox (GB) & Direct Drive (DD)



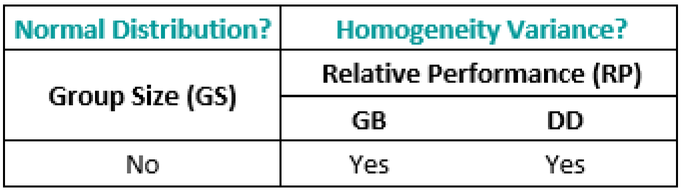
	N = 5		N = 10		N = 20		N = 30		N = 40	
	GB	DD	GB	DD	GB	DD	GB	DD	GB	DD
Levene's Statistic	0,927	1,257	1,028	2,276	0,175	0,58	0,512	0,234	0,932	0,717
df1	24	24	12	12	5	5	3	3	2	2
Sig.	0,566	0,213	0,428	0,012	0,971	0,715	0,675	0,872	0,396	0,49
Conclusion	Both DV's Sig. Value > 0.05; thus homogeneity of variance is met		Only GB has p > 0.05, and thus homogeneity of variance is met. This does not apply for DD with a p < 0.05		Both DV's Sig. Value > 0.05; thus homogeneity of variance is met		Both DV's Sig. Value > 0.05; thus homogeneity of variance is met		Both DV's Sig. Value > 0.05; thus homogeneity of variance is met	

For the DV, Table 4.7 shows the results from Levene's tests. In contrast to the test results above, all groups show Sig. above 0.05 for both the RP scores, which means that homogeneity of variance is met. Yet, the only exception to this is the group of 10. Additionally, the p-value for the RP of the DD is below the threshold to accept the null hypothesis. Hence, this variable does not accede to the homogeneity of variance. Appendix C shows the total values and tables for this test.

4.7 Conclusion

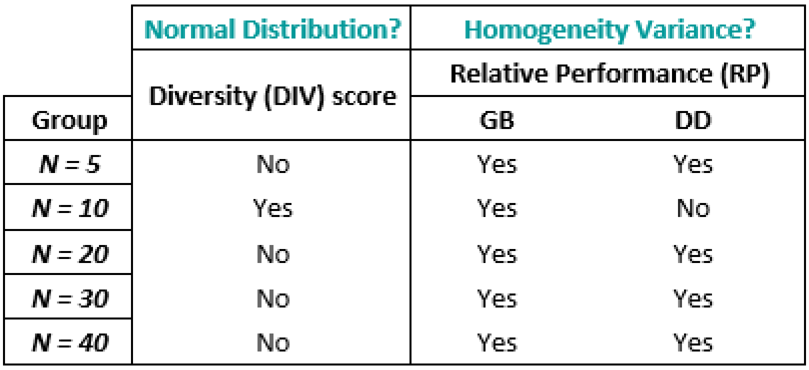
Knowing the distribution of the variables allows to select an appropriate inferential statistical technique to analyze the data. In the previous sections, an overview of several tables was provided. However, two more tables are provided summarizing the findings of these sections. From Table 4.8 and Table 4.9, the IVs do not show normal distribution. Hence, parametrical tests are not well suited in both cases.

Table 4.8: Summary findings – Group Size (GS) & Relative Performance (RP) for Gearbox (GB) & Direct Drive (DD)



Normal Distribution?	Homogeneity Variance?	
Group Size (GS)	Relative Performance (RP)	
	GB	DD
No	Yes	Yes

Table 4.9: Summary findings – Diversity (DIV) score & Relative Performance (RP) for Gearbox (GB) & Direct Drive (DD)



	Normal Distribution?	Homogeneity Variance?	
	Diversity (DIV) score	Relative Performance (RP)	
Group		GB	DD
N = 5	No	Yes	Yes
N = 10	Yes	Yes	No
N = 20	No	Yes	Yes
N = 30	No	Yes	Yes
N = 40	No	Yes	Yes

Although the IVs did not show a normal distribution, both the non- and parametrical tests were applied in this research. One drawback of this approach is that the chance of a false positive result increases when utilizing a test that assumes normality, like with the ANOVA test. Fortunately, ANOVA seems not to be very susceptible to modest divergence from normality. Various studies used a variation of non-normal distributions and concluded that the false positive rate is not affected much if the notion of normality is not satisfied (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996).

Based on the kind of variables, hypotheses, and reasoning discussed in this section, the variables were analyzed using a non-parametric test and a parametric test. Because in both cases, there were more than two groups that required comparison, hence the One-Way ANOVA test (OWAT) is applied and the non-parametric substitute dubbed the Kruskal-Wallis test (KWT).

Chapter 5

Analysis & Results

Chapter overview

This chapter investigates the relationship between the variables (DIV, GS, and RP) through inferential statistical analysis. Specifically, to find the differences between the contrived groups for the WTTs based on their GS and DIV scores, the OWAT and KWT are applied. In addition, to test for the moderation effect, the linear regression technique is applied.

- *Section 5.1 is about testing the IV 'GS' on the two DV 'RP.' Section 5.2 is similar to the previous section, but with one difference. Namely, 'GS' is replaced by the variable 'DIV-score.' Section 5.3 tests the moderation effect and discusses the process leading up to the results. Finally, in Section 5.4, the crowd results are compared with the prediction done by the expert pool.*

5.1 Group Size & (Relative) Performance

The OWAT and KWT are applied to test the hypothesis that an increase in the size of a group of people would improve their RP until the OGS, after which the RP decreases. The predictor variable GS–categorized as N = 5, 10, 15, 20, 30, and 40 and the outcome variable RP–recorded as the distance between the performance of the crowd and experts. The results are shown in Table 5.1 and Table 5.2, where Appendix E shows an overview of all the information discussed in this section.

From Table 5.1, no significant difference between the contrived groups for the DD technology was found. However, the differences between the contrived groups for the GB technology were significant, requiring a posthoc test to locate these differences for the GB group.

Table 5.1: One-Way ANOVA test (OWAT) for Group Size (GS) & Relative Performance (RP)

	F	df	Sig.
Gearbox	2,312	5	0,048
DirectDrive	0,562	5	0,729

The first posthoc test, dubbed Turkey HSD, did not significantly differ between the contrived groups. Hence, another posthoc test was necessary. Thus, the Least Significant Difference (LSD) test was selected. Table 5.2 shows the location of the differences for the GB groups.

Table 5.2: posthoc test Least Significant Differences (LSD) for Relative Performance (RP) of Gearbox (GB) Technology

	Comparison	Diff. Mean	Sig.
N = 10	N = 15	0,0231	0,016
	N = 30	0,0211	0,014
	N = 40	0,0185	0,026
N = 15	N = 20	-0,0166	0,038
N = 20	N = 30	0,0146	0,031

For the group of 10 individuals, it was concluded that when the GS increases (i.e., N = 15, 30, and 40), the RP of the group decreases. In other words, a larger group of sizes 15, 30, and 40 will perform significantly better than a group of 10. The RP decreases respectively with 0.0231 (Sig. 0.016), 0.0211 (Sig. 0.014), 0.0185 (Sig. 0.026). In contrast, the group of 15 shows that by increasing the size of the group from 15 to 20 members, the RP of the larger group increases by 0.0166 (Sig. 0.038), which means that their PERF decreases. The last difference between groups was found between the group of 20 and 30 individuals. Here, the results indicate that if the group size increases from 20 until 30, the RP will decrease by 0.0146 (Sig. 0.031).

Figure 5.1 shows the means for all the contrived group sizes and their PERF. The x-axis represents the six different treatment groups based on size. The y-axis represents the RP in comparison to the performance of experts. The lower this value, the closer the PERF of the crowd match that of the experts. In other words, their PERF increases.

After visual inspection, it can be concluded that the group of 15 members performed the best. A similar conclusion can be made for the PERF of the DD, although this was insignificant.

Table 5.3 shows the results after the application of the KWT, which is a non-parametric test. The results of the KWT are like the OWAT previously discussed. No differences were found for the DD between the contrived groups. This means that the null hypothesis has to be retained, however, this hypothesis is rejected for the GB group because differences between the groups were significant. To know where these differences lie, a series of t-tests are performed on each pair of groups. This method is also known as the Bonferroni Corrections (BC). This method was selected because it lowers the chance of committing a Type I error, which increases when conducting multiple analyses on the same dependent variable. Table 5.4 provides an overview of the results.

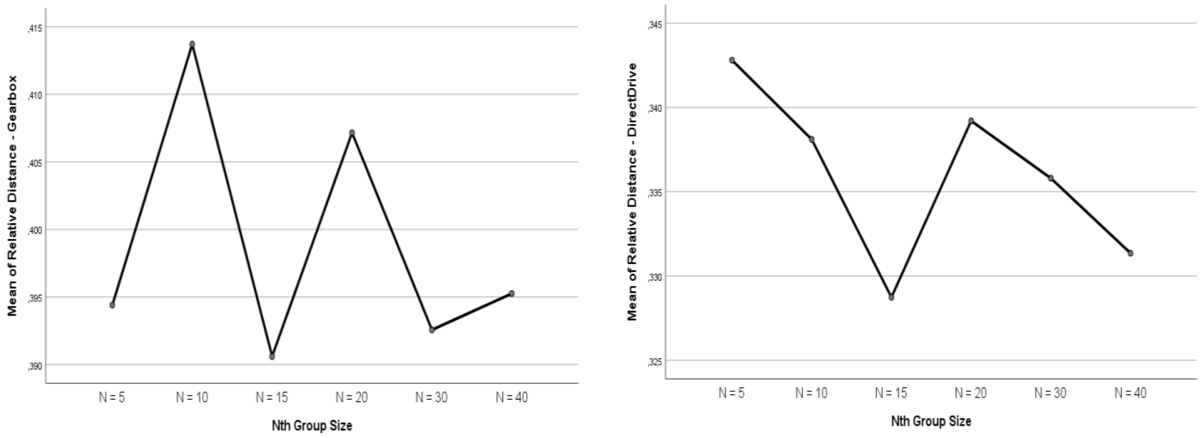


Figure 5.1: Means plot based on Nth Group Size (GS) & Relative Performance (RP) for Gearbox (GB) & Direct Drive (DD)

The data is only shown for the GS of 30, whereas comparison to the other group sizes did not lead to significant adjusted scores. This test indicated differences between the GS 30 and group sizes of 10 and 20 at p = 0.078. This value is higher than the significance value of p = 0.012 for the GB group, which indicates that the chance of a Type I error is small. Thus, the chance for a false positive (rejecting the null hypothesis when you should not) is insignificant.

Table 5.3: Kruskal-Wallis test (KWT) for Group Size (GS) & Relative Performance (RP)

	Test Statistic	df	Sig.
Gearbox	14,684	5	0,012
DirectDrive	1,844	5	0,870

Table 5.4: Post-Hoc test Bonferroni Corrections (BC) for Relative Performance (RP) of Gearbox (GB) technology

	Comparison	Test Statistic	Adj. Sig.
N = 30	N = 20	28,058	0,078
	N = 10	38,133	0,04

5.1.1 Conclusion

In conclusion, both the OWAT and KWT showed that the differences between GS and their PERF for the GB technology are significant (p << .05). With OWAT showing F = 2.312 (df = 5) at a Sig. of 0.048 and the KWT indicating a test statistic = 14.684 (df = 5) with Sig. of 0.012.

OWAT results indicate that the group of 15 had a lower RP (i.e., performance closer to that of experts) than the group of 10 by 0.0231 points with Sig. of 0.016. The group of 30 and 40 also had a lower RP when compared to the group of 10 people. Here, the differences in RP were respectively 0.0211 (Sig. 0.014) and 0.0185 (Sig. 0.026). In addition, the group of 15 and 30 also performed better than the group with 20 individuals with a respective difference of -0.017 (Sig. 0.038) and 0.0146 (Sig. 0.031).

The KWT results indicate that the group of 10 had the lowest RP compared to the group of 30 by 38.133 points with Sig. of 0.04. The same applies for the group of 20, which showed a difference of 28.058 at Sig. 0.078.

The OWAT (partially) confirms the initial hypothesis that a U-relationship exists between GS and RP in the GB case, where the contrived group of 15 indicates the optimal size. In addition, it was mentioned ‘partially’ because the findings may indicate something else as well. Namely, the RP of the group of 30 is close to that of 15 individuals, indicating that more than one optimum exists.

The results of the KWT however contradict the hypothesis and the previously mentioned results from the OWAT. Whereas the OWAT indicated that a GS of 30 performs significantly better than those of 10 and 20, the KWT shows the exact opposite. It shows that the GS of 10 and 20 outperforms the GS of 30 and that the OGS should be around 10 individuals. When combined, this test indicates a relatively linear instead of curvilinear relationship. Hence, this test would refute the initial hypothesis that a U-relationship exist between GS and RP.

5.2 Diversity & (Relative) Performance

This section describes how the OWAT and KWT test the hypothesis that a more diverse group will perform better than a less diverse group. The predictor variable is DIV-score–recorded from 0 to 1. The outcome variable is RP–recorded as the distance between the performance of the crowd and experts. Tables 5.5 and Table 5.6 show the results from the OWAT and KWT. A complete overview of all the information discussed in this section is shown in Appendix F.

Based on Table 5.5, no significant differences for all the contrived groups per alternative technology was found. There were no differences between the subgroups within each Nth group, therefore the null hypothesis is accepted.

Table 5.5: One-Way ANOVA test (OWAT) for Diversity (DIV) score & Relative Performance (RP)

	GearBox			DirectDrive		
Group	**F**	**df**	**Sig.**	**F**	**df**	**Sig.**
N = 5	1,48	24	0,09	1,58	24	0,06
N = 10	0,49	12	0,92	1,36	12	0,20
N = 20	0,65	5	0,66	1,06	5	0,39
N = 30	1,12	3	0,34	0,40	3	0,75
N = 40	0,91	2	0,40	0,63	2	0,54

Based on Table 5.6, no significant differences for all the contrived groups per alternative technology was found. There were no differences between the subgroups within each Nth group, therefore the null hypothesis is accepted.

Table 5.6: Kruskal-Wallis test (KWT) for Diversity (DIV) score & Relative Performance (RP)

	GearBox			DirectDrive		
Group	**F**	**df**	**Sig.**	**F**	**df**	**Sig.**
N = 5	33,72	24	0,09	34,83	24	0,07
N = 10	7,28	12	0,84	14,90	12	0,25
N = 20	3,42	5	0,64	5,20	5	0,39
N = 30	3,50	3	0,32	1,28	3	0,73
N = 40	1,52	2	0,47	1,33	2	0,52

However, one disadvantage when using simplified random sampling is that it is not completely random because it has a systematic way of assigning these random numbers. Hence, another technique was required to estimate DIV and RP's sampling distribution by resampling with replacement from the original sample to resolve this limitation. In other words, to get a more precise understanding of what is likely to exist in the population, thousands of alternate versions of the data set needed to be created. Hence, for the OWAT and KWT, respectively, the Bootstrapping (BOOT) and the Monte Carlo (MC) method were applied. Both exact tests enable to obtain an accurate significance level without depending on assumptions that were (not) met by the data. The resampling frequency was set at 1000 for Boot and 10000 for MC, resulting in Table 5.7, 5.8, and Table 5.9.

Table 5.7: Bootstrapping for Diversity (DIV) score & Relative Performance (RP)

	GearBox			DirectDrive		
Group	**F**	**df**	**Sig.**	**F**	**df**	**Sig.**
N = 5	1,48	24	0,09	1,58	24	0,06
N = 10	0,49	12	0,92	1,36	12	0,20
N = 20	0,65	5	0,66	1,06	5	0,39
N = 30	1,12	3	0,34	0,40	3	0,75
N = 40	0,91	2	0,40	0,63	2	0,54

Table 5.8: Monte Carlo for Diversity (DIV) score & Relative Performance (RP) for the Gearbox technology

					Monte Carlo Sig.		
	Group	**KWT - H**	**df**	**Asymp. Sig.**	***Sig.***	***Lower Bound***	***Upper Bound***
Gearbox	N = 5	26,4	17	0,07	0,06	0,05	0,06
	N = 10	6,2	10	0,80	0,81	0,81	0,82
	N = 20	3,3	4	0,51	0,51	0,50	0,52
	N = 30	1,2	2	0,55	0,55	0,54	0,56
	N = 40	0,0	1	0,84	0,84	0,84	0,85
						95% Confidence Interval	

Table 5.9: Monte Carlo for Diversity (DIV) score & Relative Performance (RP) for the DirectDrive technology

					Monte Carlo Sig.		
	Group	**KWT - H**	**df**	**Asymp. Sig.**	***Sig.***	***Lower Bound***	***Upper Bound***
DirectDrive	N = 5	19,2	17	0,32	0,31	0,30	0,32
	N = 10	10,8	10	0,37	0,38	0,37	0,39
	N = 20	4,1	4	0,39	0,39	0,39	0,41
	N = 30	1,2	2	0,55	0,54	0,54	0,56
	N = 40	0,3	1	0,57	0,56	0,56	0,58
						95% Confidence Interval	

Table 5.7 showed no significant values for all the contrived groups per alternative technology. There were no differences between the subgroups within each Nth group. Therefore, the null hypothesis is accepted. Also, no differences were found between the results from this table and Table 5.5.

As for Table 5.8 and Table 5.9, based on the asymptotic significance and the MC significance, which exceeded the p-value of 0.05. Hence, it can be concluded that there are no significant values for all the contrived groups per alternative technology, as was the case with the results for the KWT. Specifically, no differences for the subgroups within each group were found. Hence, the null hypothesis is accepted.

5.2.1 Conclusion

In conclusion, the OWAT and KWT showed that the differences between the DIV score and their RP for the GB technology are insignificant (p >> .05) for all the contrived groups independent of size. Further, for the OWAT and KWT, respectively, the BOOT and the MC method were applied to resolve the (systematic) limitation of simplified random sampling. Both techniques indicated no significant differences between the DIV score and the RP for every technology for the subgroups within each group. To conclude, all the results contradict the initial hypothesis that a more diverse group of individuals would perform better.

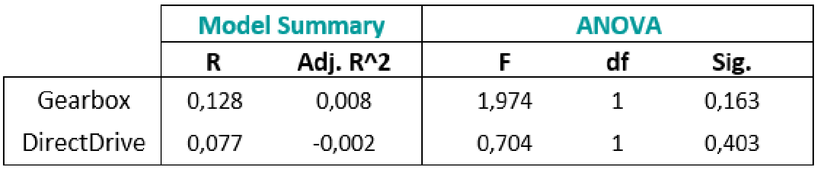
5.3 Moderation Effect

This section applied the linear regression approach to test the hypothesis that the relationship between GS and RP is positively moderated by how diverse the crowd is. The predictor variable GS–categorized as N = 5, 10, 15, 20, 30, and 40. The outcome variable RP–recorded as the distance between the performance of the crowd and experts. Table 5.10 and Table 5.11 show a brief overview of the results, where a complete overview of all the information discussed in this section is presented in Appendix G.

Before analyzing the moderation effect, a linear regression was performed for every IV individually and their corresponding DV. Table 5.10 shows the results for the regression of crowds size, and Table 5.11 shows the results of the DIV score.

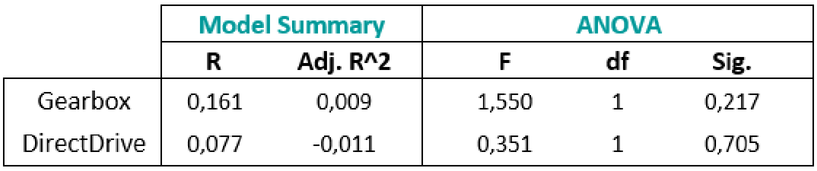
For GS, the model respectively only explains 12.8% and 7.7% of the variance in RP. As for the variable DIV score, the model only explains 16.1% and 7.7% of the variance in RP. From both the regressions, it can be concluded that the prediction capabilities are relatively poor. In addition, the results from OWAT indicate that there is no significant effect between the size of the group and RP. A similar conclusion was determined for the DIV score and RP.

Table 5.10: Linear Regression for Group Size (GS) & Relative Performance (RP)



	Model Summary		ANOVA		
	R	Adj. R^2	F	df	Sig.
Gearbox	0,128	0,008	1,974	1	0,163
DirectDrive	0,077	-0,002	0,704	1	0,403

Table 5.11: Linear Regression for Diversity (DIV) score & Relative Performance (RP)



	Model Summary		ANOVA		
	R	Adj. R^2	F	df	Sig.
Gearbox	0,161	0,009	1,550	1	0,217
DirectDrive	0,077	-0,011	0,351	1	0,705

Next, both IVs were included, requiring to perform linear regression with collinearity diagnoses to check for multicollinearity. The results are shown in Table 5.12 and Table 5.13. Once again, the model indicates poor predictabilities. Namely, 16.1% and 7.7% respectively for each WTT. Also, in both cases, the OWAT shows no significant effect between the IVs and DVs. In addition, for collinearity, the tests indicate VIF to be below the threshold value of 10. When looking into the t-coefficient and its significance value, no significant effect was measured for both cases.

Table 5.12: Linear Regression for Group Size (GS), Diversity (DIV) score & Relative Performance (RP)

	Model Summary		ANOVA		
	R	Adj. R^2	F	df	Sig.
Gearbox	*0,161*	*-0,009*	*1,55*	*2*	*0,217*
DirectDrive	*0,077*	*-0,011*	*0,351*	*2*	*0,705*

Table 5.13: Collinearity diagnoses for Group Size (GS), Diversity (DIV) score & Relative Performance (RP)

	N - group (IV)			Diversity - Score (IV)		
	VIF	t	Sig.	VIF	t	Sig.
Gearbox	*1,168*	*-1,703*	*0,091*	*1,168*	*1,06*	*0,291*
DirectDrive	*1,168*	*-0,798*	*0,426*	*1,168*	*0,066*	*0,947*

Until now, no significant effects have been found. Because of this, the focus shifted to the relationship between the moderator (DIV score) and the predictor variable (GS). To determine this relationship, the function 'Crosstabs' and 'Chi-square' statistics are used, the results of which are presented in Appendix G. Analysis of these leads to a value of the Pearson Chi-Square (PCS) of 600 with a corresponding asymptotic significance (2-sided) of 0.001, which means that the PCS is significantly below the threshold value of 0.05. Thus, it was concluded that a relationship exists between the moderator (diversity-score) and the predictor variable (crowd size).

For this (interaction effect) relationship, a new variable was computed in SPSS by multiplying both the IV, dubbed 'Interaction_1'. Then a new linear regression analysis was performed, which included this new variable. The results are shown in the two tables below.

Table 5.14: Linear Regression for Group Size (GS), Diversity (DIV) score, interaction_1 & Relative Performance (RP)

	Model Summary		ANOVA		
	R	Adj. R^2	F	df	Sig.
Gearbox	*0,168*	*0,003*	*1,125*	*3*	*0,342*
DirectDrive	*0,112*	*-0,013*	*0,488*	*3*	*0,691*

Table 5.15: Collinearity diagnoses for Group Size (GS), Diversity (DIV) score, interaction_1 & Relative Performance (RP)

	N - group (IV)			Diversity - Score (IV)			Interaction_1 (IV)		
	VIF	t	Sig.	VIF	t	Sig.	VIF	t	Sig.
Gearbox	*530,251*	*0,461*	*0,645*	*8,066*	*0,903*	*0,368*	*582,754*	*-0,542*	*0,589*
DirectDrive	*530,251*	*0,835*	*0,405*	*8,066*	*0,833*	*0,407*	*582,754*	*-0,873*	*0,384*

This model also indicates poor predictabilities, with 16.8% and 11.2% respectively for each WTT. Also, in both cases, the OWAT shows no significant effect between the IVs and DVs. In addition, for collinearity, the tests indicate VIF to be above the threshold value of 10. Exceeding this value is an issue that will be solved at a later stage. But first, when looking into the t-coefficient and its significance value, no significant effect was measured in both cases. Lastly, no significant effect for the moderator (interaction_1) variable was measured. Hence, it does not interfere with the relationship between predictor and dependent variables, which is a good thing.

As stated before, the VIF value forms an issue. The solution to this problem is by using a standardized version of the moderator and predictor variable. Resulting in two standardized variables, 'Z-N-Group' and 'Z-Div-Score.' After computing 'Crosstabs,' results were significant. Thus, a new variable is computed by multiplying the two standardized variables, then named the 'Interaction_2' variable.

A new linear regression analysis is performed, including this new variable. The results are shown in Table 5.16 and Table 5.17. The results of the analysis are very similar to the results presented earlier, as is shown in Table 5.14 and Table 5.15. But, in Table 5.16 and Table 5.17, the VIF values are now below 10. Thus, do not form an issue anymore. However, no significant values for all three independent variables were found despite the adjustments.

Table 5.16: Linear Regression for standardized variables (i.e., Group Size (GS), Diversity (DIV) score, interaction_2 & Relative Performance (RP)

	Model Summary		ANOVA		
	R	Adj. R^2	F	df	Sig.
Gearbox	*0,168*	*0,003*	*1,125*	*3*	*0,342*
DirectDrive	*0,112*	*-0,013*	*0,488*	*3*	*0,691*

Table 5.17: Collinearity diagnoses for standardized variables (i.e., Group Size (GS), Diversity (DIV) score, interaction_2 & Relative Performance (RP)

	N - group (IV)			Diversity - Score (IV)			Interaction_2 (IV)		
	VIF	t	Sig.	VIF	t	Sig.	VIF	t	Sig.
Gearbox	2,501	0,764	0,446	2,164	0,409	0,684	2,457	-0,542	0,589
DirectDrive	2,501	0,093	0,926	2,164	-0,544	0,587	2,457	-0,873	0,384

The analysis of the moderation effect is the final step. The model works correctly based on a visual inspection. Therefore, the linear regression of the previous step is repeated but independently for every IV. The results are shown in Figure 5.2.

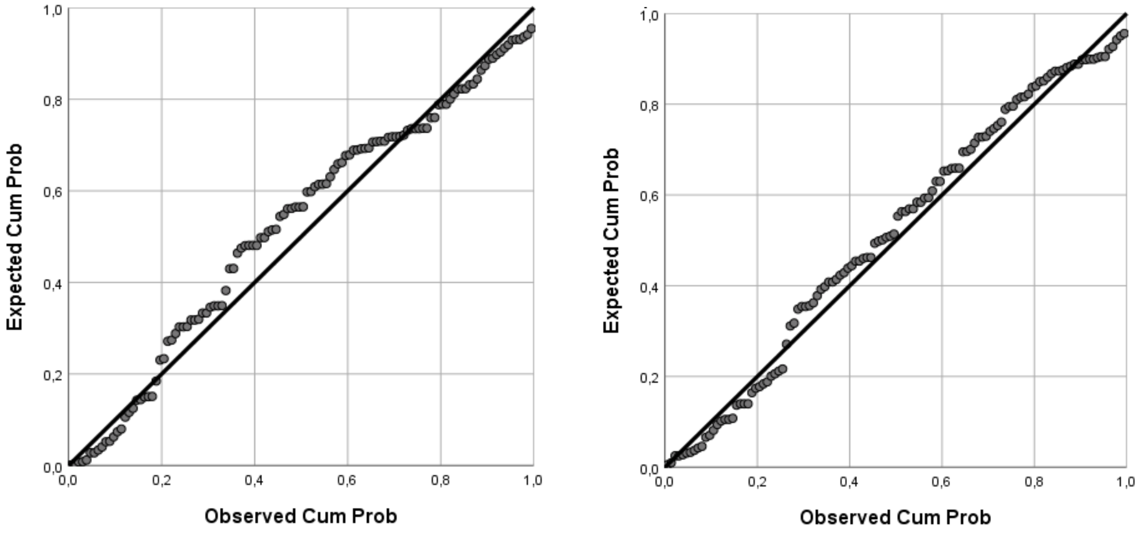


Figure 5.2: Normal P-P plot of Regression standardized residual dependent variable: Relative Performance (RP) Gearbox (GB) & Direct Drive (DD)

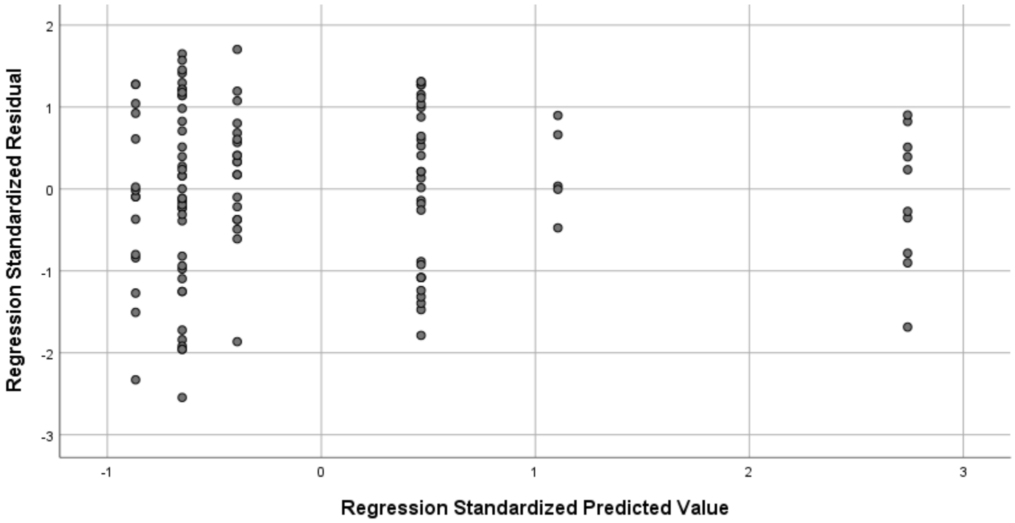


Figure 5.3: Scatterplot dependent variable – Relative Performance (RP) Direct Drive (DD)

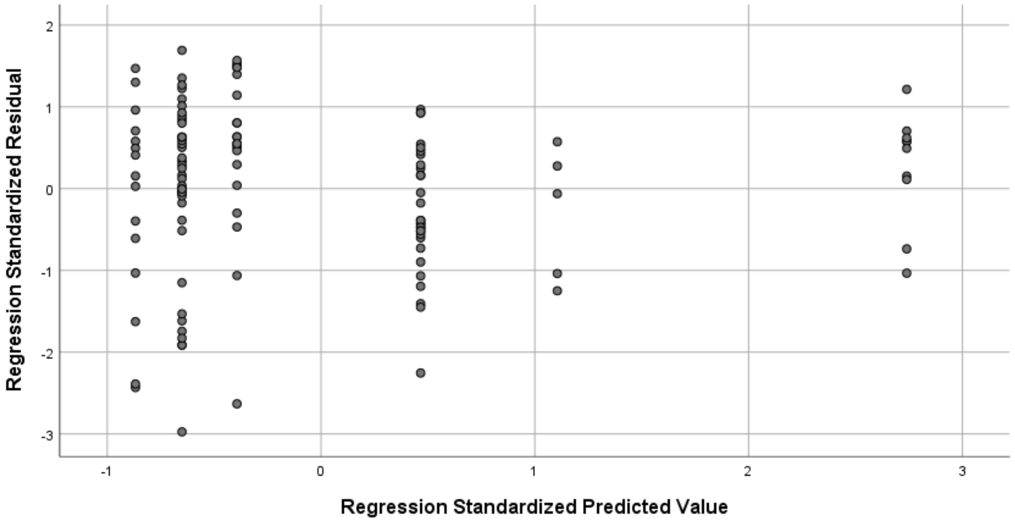


Figure 5.4: Scatterplot dependent variable – Relative Performance (RP) Gearbox (GB)

In both cases, all the variables are spread evenly and indicate normality when looking at the scatterplots for both technologies. Because they are all distributed vertically with different points along the x-axis, it was concluded that the model, even though it only predicts 16% of the variance. It can still predict the variance of RP this well for this percentage. The last check is to verify if the model is correct based on the P-P plots. All points show minor deviations from a straight line, even though there are no significant results.

5.3.1 Conclusion

The linear regression was used multiple times. In the first case, this was done separately for every IV. For GS and DIV score, the results for the GB and DD were respectively F = 1.974 (df = 1) at Sig. 0.163, F = 0.704 (df = 1) at Sig. 0.403 and F = 1.550 (df = 1) at Sig. 0.207 and F = 0.351 (df = 1) with Sig. 0.705. For both variables, no significant effect was measured between GS and RP and DIV score with RP.

Second, linear regression was applied to see if the moderation effect exists. The results for the GB PERF were F = 1.55 (df = 2) at Sig. 0.217. For the DD, this was F = 0.351 (df = 2) at Sig. 0.705. Thus, it was concluded that no significant effect was measured.

Third, to find out if there exists any relationship between the moderator (DIV score) and the predictor variable (GS). The function 'Crosstabs' and 'Chi-square' statistics were applied. The value of the PCS is 600 with a corresponding asymptotic significance (2-sided) of 0.001, which means that the PCS is significantly below the threshold value of 0.05. Thus, it was concluded that a relationship exists between the moderator (DIV score) and the predictor variable (GS).

Fourth, another linear regression analysis was performed, introducing a new variable (i.e., interaction) to correct the exceeding VIF value. The results for the GB and DD respectively were F = 1.125 (df = 3) at Sig. 0.342 and F = 0.488 (df = 3) with Sig. 0.691. Thus, no significant values were found for all the three independent variables for both the technologies.

To conclude, the multiple linear regressions did not show any significant effect even after corrections for the VIF value. Thus, the results contradict the initial hypothesis that the relationship between GS and RP is positively moderated by how diverse the crowd is. However, in contrast to this, there was indeed a significant relationship between DIV and GS.

5.4 Comparing Results & Predictions

This section does not apply (non) parametric tests because this sub-research question can be answered with only descriptive statistics. This question aims to find out what the results of the CI of the crowd indicate compared to experts by comparing the results on an individual and group level with those of the expert group. The comparison was based on the survey's duration, consistency ratio (i.e., if comparisons were made consistently and if the results are reliable), selection of the best and worst criterion, and comparing the final PERF score per alternative technology.

Table 5.18 shows the statistics for the survey duration time, experts were not included in this table because this information was not collected. However, the time to complete an interview with one expert is assumed to take approximately three hours. This was based solely on experience and did not involve any post transcription or calculations. In total, twelve interviewees got interviewed, resulting in a total time of 36 hours.

When comparing the total time invested between experts and individuals (excluding preparation time), it is noticeable that the maximum time to completion for the crowd is roughly 50 minutes. Similar observations can be made for the groups of individuals, where the maximum completion time was roughly 39 minutes. In both cases, the time spent was lower than the time required to conduct expert interviews. In addition, the groups of 5, 10, 15, and 137 respondents were faster than the individual's mean. The individual completed the survey with less time than the groups of 20, 30, and 40.

Table 5.18: Survey completion time in seconds

Group size	Mean	Median	Std. Dev.	Min.	Max.
Individual	456	360	413	90	3093
N = 5	255	202	120	157	453
N = 10	330	332	163	105	619
N = 15	393	259	288	146	1062
N = 20	512	287	643	153	3093
N = 30	475	370	412	90	1885
N = 40	494	412	376	93	2318
N = 137	441	332	397	90	3093

Table 5.19: Consistency ratio of crowd, expert pool & the individual

Consistency Ratio	Statistics	Experts	Individual	N = 5	N = 10	N = 15	N = 20	N = 30	N = 40	N = 137
Ksi - Gearbox	Min.	0,112	0,031	0,048	0,065	0,031	0,035	0,044	0,032	0,031
	Max.	0,112	0,230	0,088	0,175	0,172	0,205	0,193	0,230	0,230
	Mean	***0,112***	***0,105***	***0,066***	***0,109***	***0,097***	***0,106***	***0,112***	***0,107***	***0,104***
	Std. Dev.	.	0,044	0,017	0,042	0,046	0,039	0,048	0,045	0,045
Ksi - DirectDrive	Min.	0,112	0,012	0,030	0,050	0,028	0,012	0,019	0,017	0,012
	Max.	0,112	0,271	0,113	0,157	0,192	0,209	0,271	0,195	0,271
	Mean	***0,112***	***0,106***	***0,079***	***0,104***	***0,109***	***0,110***	***0,113***	***0,102***	***0,107***
	Std. Dev.	.	0,051	0,034	0,033	0,054	0,056	0,059	0,049	0,052

In Table 5.19, the consistency ratios of the expert group and crowd (i.e., individual and groups) are shown. To calculate this, all the steps to test for consistency as proposed by (Rezaei, 2016) were applied. The largest value for the expert group was 0.112, for the individuals 0.106, and the Nth groups 0.113. Thus, the Nth groups, experts, and individual results proved to be consistent, which means that the outcomes are very reliable. However, the mean of the individuals proved more reliable than that of the expert pool. Similarly, all Nth groups except the group of 30 showed a lower consistency ratio than that of the experts. For GB, the groups of 5, 15, and 137 have a lower value than the individuals. For DD, the groups of 5, 10, and 40 had more reliable results than the individuals.

Table 5.20 and Table 5.21 illustrate the results for selecting the best and worst determinants. Each number in the table represents one of the eight criteria, the definition of which is described in the codebook in Appendix B. As for selecting the best criterion for both the WTT on an individual and group level, the majority selected number three, which is the criteria known as 'cost of energy' followed by 'reliability.' This is quite in line with the choice of the expert group. Namely, the results from the paper (van de Kaa et al., 2020) showed that an essential factor appeared to be 'cost of energy,' closely followed by 'reliability.' The worst criteria selected by the experts was 'pre-emption of scarce assets' followed by 'brand reputation and credibility.' In contrast, the individuals and groups choose 'reliability' as the worst determinant and, in some cases, 'Pricing strategy.' The former choice however is rather peculiar and contradicts their first consideration when selecting the best determinant. Despite this, it was concluded that there is indeed a difference in selecting the worst criterion.

Table 5.20: Descriptive statistics – Selection of the best criteria

Group Size	Best Criterion	Mean	Median	Std. Dev.	Min.	Max.
Individual	Gearbox	3,2	**3**	1,6	1	8
	DirectDrive	3,2	**3**	1,5	1	7
N = 5	Gearbox	3,8	**4**	1,1	2	5
	DirectDrive	3,4	**4**	1,8	1	5
N = 10	Gearbox	3,5	**4**	2,2	1	7
	DirectDrive	3,5	**3,5**	1,5	2	7
N = 15	Gearbox	3,3	**3**	1,5	1	6
	DirectDrive	2,9	**3**	1,2	1	5
N = 20	Gearbox	3,6	**3**	1,4	2	6
	DirectDrive	3,5	**4**	1,5	1	7
N = 30	Gearbox	2,9	**2,5**	1,9	1	8
	DirectDrive	3,2	**3**	1,5	1	7
N = 40	Gearbox	3,1	**3**	1,5	1	8
	DirectDrive	3,1	**3**	1,5	1	7
N = 137	Gearbox	3,2	**3**	1.6	1	8
	DirectDrive	3,3	**3**	1,4	1	7

Table 5.21: Descriptive statistics – Selection of the worst criteria

Group Size	Worst Criterion	Mean	Median	Std. Dev.	Min.	Max.
Individual	Gearbox	4,2	**4**	1,9	1	8
	DirectDrive	4,5	**4**	1,7	1	8
N = 5	Gearbox	4,2	**5**	1,6	2	6
	DirectDrive	4,6	**4**	0,9	4	6
N = 10	Gearbox	4,5	**4,5**	1,8	2	8
	DirectDrive	4,3	**4**	1,8	2	8
N = 15	Gearbox	4,5	**4**	2,1	1	8
	DirectDrive	4,1	**4**	1,8	2	7
N = 20	Gearbox	3,9	**3,5**	2,3	1	8
	DirectDrive	4,5	**5**	2,0	1	7
N = 30	Gearbox	4,3	**4**	1,7	1	7
	DirectDrive	4,3	**4**	1,3	2	7
N = 40	Gearbox	4,1	**4**	1,8	1	8
	DirectDrive	4,8	**5**	1,8	1	8
N = 137	Gearbox	4,1	**4**	1.9	1	8
	DirectDrive	4,5	**4**	1,6	1	8

Table 5.22 compares the PERF scores, which shows that the individual and all the groups independent of their size do not show any similarities with the expert pool. In general, the findings indicate that the individual and all the groups have a grade that's two-thirds of the value given by the experts independent of WTT. Hence, it was concluded that there are differences between how the crowd and the experts think the technology performs. However, some groups did manage to perform relatively better than any individual. Namely, for the GB, the groups of 5, 15, and 30 showed their difference compared to experts to be lower than the individuals. Like DD, the groups of 15 and 40 had a lower difference to the experts than to individuals.

Finally, by grading the technologies, a prediction can be made. When looking into the experts' opinion, both technologies score about evenly. Therefore, they predicted that the battle was not over yet, that both technologies have an equal probability of winning, or that there is no winner and the two WTTs co-exist on the market. Hence, it was concluded that all groups and individuals would draw a similar prediction to the experts, even though their final grades differed by two-thirds.

Table 5.22: Comparing the performance score

Performance Score	Statistics	Experts	Individual	N = 5	N = 10	N = 15	N = 20	N = 30	N = 40	N = 137
Gearbox	Min.	0,528	0,092	0,113	0,092	0,099	0,095	0,106	0,093	0,092
	Max.	0,528	0,203	0,156	0,145	0,191	0,194	0,182	0,203	0,203
	Mean	**0,528**	**0,131**	**0,134**	**0,114**	**0,137**	**0,121**	**0,135**	**0,133**	**0,130**
	Std. Dev.	.	0,024	0,019	0,016	0,029	0,024	0,018	0,025	0,125
DirectDrive	Min.	0,472	0,095	0,112	0,108	0,107	0,095	0,103	0,097	0,095
	Max.	0,472	0,204	0,147	0,174	0,199	0,186	0,182	0,204	0,204
	Mean	**0,472**	**0,138**	**0,129**	**0,134**	**0,143**	**0,133**	**0,136**	**0,141**	**0,137**
	Std. Dev.	.	0,135	0,014	0,021	0,027	0,020	0,026	0,029	0,135

5.4.1 Conclusion

In general, the time spent conducting the BWM questionnaire for the crowd was substantially lower than performing experts interviews. In only two cases (i.e., N = 5 & 15), the group was faster than the individual.

Secondly, all the results proved to be consistent and thus highly reliable. The individuals and all other groups proved to have more reliable results than experts, except for the group of 30. When comparing on an individual and group level, it shows that the groups of 5 and 15 have a lower value than the individuals for the GB. As for the DD WTT, groups 5, 10, and 40 had more reliable results than individuals.

Third, most individuals and contrived groups choose the number three when selecting the best criterion for both the WTT. Number three is the criteria known as 'cost of energy' and was followed by number four, known as 'reliability.' This is quite in line with the choice of the expert group. In contrast, the individuals and groups choose 'reliability' as the worst determinant, and in some cases, 'Pricing strategy.' However, the former choice is rather peculiar and contradicts their first consideration when selecting the best determinant. Hence, this part does not line up with the selection done by the experts. Despite this, it was concluded that there is indeed a difference in selecting the worst criterion.

Fourth, for the comparison based on the PERF score, the findings indicate that the crowd grades the WTT two-thirds lower than the experts'. Hence, it was concluded that no individual nor group performed in similar ways like the expert pool. Furthermore, for the GB, the groups of 5, 15, and 30 showed differences compared to experts, which was lower when compared to individuals. Like the DD, the groups of 15 and 40 had a lower difference than the individual compared to experts.

Fifth, it can be concluded that all the groups and individuals would draw a similar prediction like in the case with experts even though their final grade differs by two-third. Because the difference in PERF score for the WTTs was less than 0.058. This number was based on the maximum difference between the experts' performance scores for the two WTTs.

Chapter 6

Discussion: Significance of the findings

Chapter overview

This chapter delves into the meaning, importance, and relevance of the results obtained in this research. The focus is on explaining and evaluating the results and showing how this relates to the literature review and research questions and making an argument supporting the overall conclusion in the next chapter.

- *Section 6.1 significant findings are concisely summarized into a clear statement of the overall results, answering the main research question. Section 6.2 discusses the meaning of the results, where the overall aim is to show the reader precisely what this research contributed and its importance. In section 6.3, the limitations of this research are discussed and how these limitations influenced the results. Lastly, in section 6.4, recommendations for future research are shown, based on the results in the previous sections.*

6.1 Key findings

In section 1.3, the research problem was identified. An issue hampered the representativity and specificity of experts' predictions, often due to the paucity of available, knowledgeable, and inclined experts to participate in the BWM questionnaires. Hence, this research focused on understanding, testing, and examining if the CI of a crowd could perform similarly to experts. From this point, the main research question guiding this research was:

Main Research Question

How does the Collective Intelligence perform in comparison to experts when predicting technologies?

This question is answered by the results for the sub-research questions, as was described in Chapter 5. How CI operates depends on various factors, such as survey completion time, consistency ratio, selection of best and worst criteria, PERF grade, and the final prediction. In addition, the influence of the variables 'DIV score' and 'GS' on the PERF of the crowd was investigated.

Based on the factors listed above, the CI of a crowd was successful to some degree. For instance, the time spent conducting the survey was substantially lower than interviewing experts. Also, the individuals and all groups (excluding the group of 30) proved to have more reliable results than experts. As for selecting the best criterion, both individuals and groups choose a similar answer to the experts. However, both did not manage to perform like experts when choosing the worst determinant. This choice contradicts their second choice when selecting the best factor because they are the same. As for the comparison based on the PERF, the crowd graded the WTTs two-third lower than the value given by experts. Even though no individual or group performed similarly to the experts, all the groups and individuals would draw a similar prediction to the experts.

Furthermore, differences between the DIV score and RP for both WTT were insignificant. Hence, the initial hypothesis that a more diverse group of individuals would perform better was refuted. As for the moderation effect, a significant relationship between DIV and GS was found. However, the multiple linear regressions did not show any significant moderation effect even after standardizing the variables. Hence, the initial hypothesis that the relationship between GS and RP is positively moderated by how diverse the crowds were was rejected. Lastly, for both cases it was not possible to state how DIV and GS influence the PERF of the crowd.

For the variable GS, significant differences were found between the contrived groups in the case of the GB technology. Nevertheless, the findings for both the tests contradict each other and the hypothesis. Namely, OWAT provides proof that there is indeed a U-relationship between GS and RP. The OGS contained 15 people. In contrast, the KWT indicates a relatively linear relationship, where the groups of 10 and 20 performed significantly better than the group of 30. However, OWAT showed that the group of 30 performed significantly better than the group of 10 and 20 people. In addition, it was concluded the OGS to be at 10 and not 15 individuals.

To conclude, this research investigated the effects that GS and DIV could have on the PERF of the collective. Only the findings for the variable GS were deemed significant in the case of the GB WTT. However, both the OWAT and KWT contradict each other but also the hypothesis. Consequently, this research was not able to conclude how these variables affect the PERF of the crowd. Nevertheless, this research must conclude that the crowd performs differently than experts when predicting standard battles. The PERF score of the crowd was two-thirds lower than that of the experts. In some aspects however the crowd performed in similar or better ways. Due to only testing one case, it limited our insights if this happened due to chance or not. In addition, the main goal of this study was to investigate if the crowd could come up with the exact prediction. This does not entail that only the conclusion should be the same, but the process (e.g., selection criteria and grading WTT) towards this conclusion should also be similar. However, this was not the case. Hence, the crowd performed differently than experts when predicting the outcome of the WTT battle. In addition, the consistency ratio remains a matter of doubt since most of the results were initially unreliable. Therefore, it was necessary to make (logical) corrections to obtain reliable results. This makes the comparison based on the consistency ratio rickety.

6.2 Interpretations & Implications

Now that the entire research has been presented, the specific implications are broken down into more detail. This discussion starts below.

In the literature review, many ambiguous and contradicting claims were noticed, such as how big the size of the contrived group should be to perform similar to experts, but also what constitutes DIV and if it influences the PERF of the crowd. In addition, based on the previous claims, it was also investigated to which extend the moderation effect takes place for these two characteristics.

For GS, the findings were not entirely in line with supporting this hypothesis. Namely, due to differences in results for the non- and parametric tests. Furthermore, the theory depicts that only one optimal level of GS exists, but consensus on where this optimal level would not have been reached. However, when you increase or decrease the GS from this point, the PERF of the group would decrease (Hashmi, 2005). Based on the OWAT findings for the GB technology, the findings indicate that the PERF improves significantly by increasing the group size from 10 to 15. However, when increased from 15 to 20 members, the PERF decreases significantly—indicating a U-shape relationship. Hence, strengthening the claim of (Hashmi, 2005) and challenging the claim of (Carvalho et al., 2016). Because the group of 20 did not perform slightly better than the group of 10. The results from KWT confirm this view. However, their claim that the group of 10 performed better than an individual is supported based on Table 5.19.

The results in Figure 5.1 also suggest that the group of 15 and 30 perform pretty much the same, suggesting that there may be two optimums instead of only one. Thus, it contradicts the claim of a U-shape relationship with only one optimal point. According to (Hashmi, 2005), one would expect the group of 30 and 40 to follow this decrease of PERF as was noted with the group of 20. This was not the case however, as they performed pretty much the same as the group of 15. Nevertheless, it is noticeable that the slope of this increase in size is not proportional to the PERF. It decreases by a small amount, which could indicate that two optimums exist and many more. However, the PERF slightly decreases for every increase of size by a multitude of the OGS (i.e., 15). Although this was not verified, it does provide ample indications that the claim of (Green, 2015) (i.e., slope decreases as group size increases) seems plausible. This is only the case when the size of the group is a multiple of the OGS and not when GS increases, because the results from the group of 20 contradict this claim. However, (Green, 2015) also proposed that larger groups perform better than smaller groups independent of the tasks. Yet, the results do not support this claim. Because the PERF of the group of 15 and 30 was better than the group of 40.

In contrast to the OWAT, the findings from the KWT for the GB technology indicate something different by contradicting the claim of (Hashmi, 2005) because the results suggest a linear relationship between GS and their PERF instead of the U-shape relationship. Furthermore, this may indicate that the claim of (Green, 2015) may be valid because the PERF of the crowd decreases when GS increases.

The results of the OWAT and KWT indicate different but closely related OGS. For the former this was a group of 15 and for the latter a group of 10. Both test results give indications that support the claim of (Carvalho et al., 2016), which asserted that the optimal size is independent of the task and motivation and that OGS was around 10 to 11 people. However, (S. Krause et al., 2011) observed an optimal point much higher, around 25 agents for a medium-sized task. Nevertheless, the findings do not support this. Because the performance of the group of 20 is much lower than that of 15 or even 10. To conclude, this research could not locate the OGS due to contradicting test results and statements about the OGS.

As for the theory on DIV and PERF, both the BOOT and MC results do not indicate a relationship. Hence, the initial hypothesis that a more diverse group of individuals would perform better is refuted. These findings simultaneously support and contradict different claims. Namely, the theory of (S. Krause et al., 2011; Nguyen et al., 2018; Surowiecki, 2005) underlined the importance of DIV in respect to the PERF. Especially for Nguyen and Surowiecki, who indicated that DIV is the most important criteria for a group of people to act intelligently. The findings put their claim to the test because the relationship was insignificant. In contrast, the results give indications that the claim of (Reynolds et al., 2017) may be valid. They concluded that there was no correlation between DIV and PERF but that there was indeed a strong correlation between cognitive DIV and PERF. Yet, cognitive DIV was not tested because it was hard to detect and measure, thus this assertion is left unaffected. However, there is ample support in their claim that DIV is indeed not the magic bullet.

Another implication was found after multiple linear regressions but with no significant effect. In other words, the initial hypothesis that the relationship between GS and PERF is moderated by how diverse the crowd is weakened. Nevertheless, the findings indicate that there was indeed a significant relationship between DIV and GS. Although a significant relationship between GS and DIV was established, the other findings did not support the claims of (Oliver et al., 1988) and (L. Robert et al., 2015). The reasoning behind this stems from the previous interpretation of DIV and cognitive DIV. Namely, that there is no relationship between DIV and PERF. Based on this information, it would make sense that there would be no moderation effect.

Next, the predictions made by the crowd and experts were compared. However, no other research was found that studied a similar subject like this research. Thus, no comparison could be made between studies. Instead, various statements are compared that were collected during the literature review. For example, indicated that the crowd could perform similarly to experts (Larrick et al., 2012; Mollick et al., 2016; Ray, 2006; Sunstein, 2008; Tetlock, 2017). While some claimed the crowd could outperform the experts (Adomavicius et al., 2005; Budescu et al., 2015; Larrick et al., 2012; Lorenz et al., 2011). While additional researchers proved that advice from the CI reduces the variance of recommendations to its mean, enabling superior decision-making compared to that of an individual (Chiu et al., 2014; Kittur et al., 2008; Kozinets et al., 2008; Lorenz et al., 2011).

Various statements are challenged based on the findings, while some are partially supported. Specifically, if the crowd could perform similar to experts by comparing the consistency ratio, selection of the best determinant, and the final prediction drawn from the performance score. However, this was not the case when selecting the worst criteria and grading the technology where individuals and groups were approximately two-thirds lower than the grade given by experts. Thus, it can be concluded that there is support in the claim of (Larrick et al., 2012; Mollick et al., 2016; Ray, 2006; Sunstein, 2008; Tetlock, 2017) but only in certain areas. Furthermore, the only factor in which the crowd exceeded the experts was the time required to complete the survey versus the time required to conduct expert interviews. In addition to this, the motivation of the crowd and experts differs vastly. Namely, the expert's motivation is to share their knowledge and time freely, in contrast to people completing the survey whose primary motivations are likely monetary incentives. Hence, it can be argued that the crowd may have a higher incentive to complete the survey faster than an expert would. In other words, the claims of (Adomavicius et al., 2005; Budescu et al., 2015; Larrick et al., 2012; Lorenz et al., 2011) are refuted.

Finally, this research also compared the results on an individual as well as group level. However, they did not show any superior decision-making for the group compared to the individual, based on considering the survey's duration, the consistency ratio, the selection of the best and worst criterion, and the comparison of the final performance score per alternative technology. For many of these cases, the variance of the groups in comparison to the individual was very similar. The same is also true for the values of individuals and groups in each case, which suggests that the assertions by (Chiu et al., 2014; Kittur et al., 2008; Kozinets et al., 2008; Lorenz et al., 2011) are not supported.

6.3 Limitations

Generally, limitations of a study are considered those characteristics of a design or methodology that impacted or influenced the interpretation of the research findings. By acknowledging such limitations, it opens the opportunity to demonstrate new knowledge, but also to confront the assumptions taken as well as explore what is not known. The limitations of this research are discussed and acknowledged through evaluating how these influenced the findings and interpretations.

Grasping the CI / CS literature

Harvesting human wisdom has proven a promising resource for various purposes. Though in recent years, many researchers have attempted to understand CI from different domains ranging from sociology up to computer science, but with no consensus on what constitutes and affects the performance of CI. One cause for this may come from some scholars who use the notion of CS and CI interchangeably, which caused confusion because the two ideas seem similar, but still are different. Another reason may be the numerous taxonomies varying from mass collaboration to crowd wisdom and several others, which would explain the sometimes blurry and ambiguous definitions found in the literature.

This limitation made it more difficult to fully comprehend the notion of CI as it took a vast amount of time to select studies and reliable findings. In addition, prior research studies of CI in the context of standard battles and BWM were practically non-existent, which resulted in an exploratory research design instead of an explanatory one.

Preparing, testing & gathering data for Analysis

As stated in chapter 3, in this research, the control variable was 'independence,' which was the second most important condition to grasp the complete potential of CI. However, this research explained why it was 'controlled.' This was mainly since it was not possible to verify whether the individuals independently responded to the survey, but also had to do with the contradicting claims stating that the crowd should be independent or not. Nevertheless, this condition in administrating an online survey seems paradoxical because it is required to provide some background information for the respondents to help them understand the problem at hand. However, independence is described as giving your opinion based on your own experience and knowledge. This means that providing this kind of information would violate this condition, as it could steer the respondents in a particular direction. It is therefore likely that this had some kind of impact on the results. Nevertheless, let's assume independence is not met. The consequence would be that the performance of CI decreases or increases depending on your view. However, the magnitude of this impact on the performance is unknown, making it difficult to fully comprehend its impact on the findings. Still, to determine if the assumption of independence was plausible. The data needed to be rich in variance. If this was not the case, this could indicate the condition is not satisfied. In the end, the data mainly indicated normal distributions, such as the distribution of the DV's and the frequency of the survey completion time, with a relatively low meantime. This makes the assumption plausible, thus limiting the adverse effects of not complying with this condition.

Another unexpected limitation involved the consistency ratio of the respondents. This limitation was anticipated and solved by implementing some functions in Qualtrics (e.g., carrying for the selected answer) and adding colors for the best and worst criteria to reduce confusion or obliviousness. Despite these efforts, many results were not reliable. Hence, the method of solving this was deemed unsuccessful. When investigating why this was the case, it was noticeable that many respondents did not grade the criteria correctly. When comparing the best against the best criteria or the worst against the worst criterion, one would assume that both will get an identical score because there are no differences. When comparing the best and worst criteria, one would assume people select a score that indicates the most significant difference. Unfortunately, this was not the case. To correct for this, the selection of best and worst criteria has been looked into for each individual and the correctness of the grading has been verified. For the cases where the grading was incorrect, the grades were changed to a one or a nine, depending on the comparison. By doing so, the consistency ratio was now similar to that of the experts. In other words, the results were reliable and could be further analyzed.

In section 4.5.2, the characteristics of the sample were discussed. It consisted primarily of young males from the Americas that were highly educated and worked full time. In hindsight, this is not incredibly diverse, which was required to fully grasp CI's capabilities. The characteristics were limited in DIV because only one platform (i.e., MTurk from Amazon) was used to distribute the survey in combination with a technological topic. In addition, due to the GQ, 65 respondents were excluded from this study. This means that roughly 30% of the respondents had been excluded, which may have limited the DIV in the sample. The GQ was a multiple choice question with four possibilities and formulated as follows: "How much time does the earth take to revolve once around the sun?". In hindsight, the question may have been too complicated or leaning too much towards science. With this in mind, it is likely that people with higher education, especially with a background in physics, chemistry, etcetera, would be more likely to answer correctly. Looking at the excluded respondents however, the vast majority was higher educated, but it could not be established in what area. Nevertheless, they could not give the correct answer, indicating that the question was either too difficult, too specific, or that the respondents were rightfully excluded.

The survey results were used to find out whether CI of a contrived group could perform similar to experts when conducting the BWM. This research however was limited in its resources, resulting in primary and secondary data to accomplish this goal. In other words, all the data presented was gathered from the respondents. Nevertheless, the calculations for the final PERF grade given for each alternative technology could not have been calculated without the tables from (van de Kaa et al., 2020). This entails that the weights of the criteria did not follow from the subjects' data but from the experts, which could have influenced the results to some degree. In the same research however, two pools of experts were utilized. The first figured out how important each criterion was. The second group of experts graded the technologies based on these criteria without knowing the actual weights. This makes it plausible to assume that this influenced the PERF score of the respondents, but not any more or less than the second expert pool. This leaves a knowledge gap that is elaborated in the recommendations section.

It should be pointed out that human error may have influenced the results of this research. More specifically, the slips and lapses that may inadvertently occur while routinely doing more than 600 manual calculations, not even including the corrections that had been done for the consistency ratio. It was made sure to make shorter files and only work for a limited time to keep myself concentrated on reducing the risk of slips. Also, with every calculation, the respondents' ID number was compared to verify that the data belonged to this person.

Analyzing and comparing CI

The first limitation in this area is the PERF that is measured with the survey. When comparing the two RP of the WTT, it became clear that almost everyone performs better for the GB (which was graded first) than for the DD. This probably occurred for several reasons. For example, the respondents answered the same questions asked in an expert interview to allow a better comparison. Still, the results suggest that the second grading task was too much to ask because the first task was too complicated or maybe too hard to solve, and thus requiring a considerable amount of effort and resources. This left the subjects with little to no resources (e.g., mental, time) left. If this were true, it could indicate that the assumption of defining the task as one with fine granularity is incorrect, which means that the task should be defined as coarse or, in other terms, complex. In addition, the monetary incentive offered when completing the survey could have influenced the difference in PERF. For example, some respondents could have concluded after finishing the first part that the monetary incentive was too low. Nevertheless, to still get the reward, they would speed up the second part of the survey, resulting in this difference in PERF score. Even though the total time required to complete the survey is available, the time required to complete the individual tasks is not. Hence the validity of this statement could not be checked. Thus, nothing could be concluded.

This section will not elaborate on the (dis)advantages of the inferential statistics techniques used, such as OWAT and the KWT, because of the large amount of information that is available online. Limitations in other areas are highlighted. For the MC technique, the number of subjects was sufficiently large, but the range of most DIV scores was poor, affecting the significance of the MC results. Further, only one case was applied in the survey, thus verifying the results of the first case with another was not established. This did not influence the interpretation of the findings. However, deeper insights could not be collected, making it a limitation. The moderation effect was also tested for but was insignificant, thus weakening some claims found in the literature review. This probably occurred due to the research setup, which required more respondents to have sufficient data points. This was noticeable because there was plenty of data on how respondents graded the technology, but only six DIV scores represented the six groups. In combination with a limited DIV score (as mentioned before), this could have influenced the findings of the moderation effect. Lastly, the relationship between DIV and GS was established, but what this relationship entails could not be established due to limitations in resources. This provides room for future researchers, which will be further discussed in Section 6.5.

6.4 Recommendations

In this section, the suggestions for the practical implementation of further research is discussed. In this case, the recommendations for further research follow directly from the limitations section. In other words, this study provides concrete ideas for how future work could build on areas that this research could not address.

Grasping the CI / CS literature

As mentioned before, many definitions and similar concepts are considered. Incorporating a variety of interpretations was beyond the scope of this study. Hence, it has yet to emerge. In addition, more research is required to establish the difference between the two concepts (i.e., CI and CS) and avoid confusion in the future. Furthermore, because prior research was non-existent in the context of MCDM and standard battles, this research was confined to an exploratory research design. Thus, future studies should build on this prior research to conduct explanatory research.

Preparing, testing & gathering data for Analysis

This study did not verify the control variable 'independence.' However, in future studies, researchers can verify this condition by simply asking the respondents whether they have completed the survey independently. In addition, solving the paradox between independence and providing background information for the task is complex. Thus, it is recommended to keep the information provided to respondents similar to the information provided to the expert pool to minimize the difference between the two, hence minimizing the influence when comparing the two groups. Furthermore, if the condition of independence is not satisfied, the consequence on the PERF of CI decreases or increases depending on your view. Yet, how much this would affect the PERF is still unknown. Therefore, future research is needed to establish the effects that independence has on the PERF of the crowd.

As for the consistency ratio, the respondents' results were not reliable for the majority of respondents. The reason why this happened is yet unknown, although the study of (Nakatsu et al., 2014) could provide some insights. The distributed task for the crowd was unstructured. In other words, the task had no clearly defined solution. This situation probably occurred because of ambiguous boundaries of knowledge domains. Hence, the recommendation for future studies is to investigate multiple cases to see if this occurs all over again. If so, the next logical step would be to find out why this transpired.

The characteristics of the sample were limited in DIV, which likely followed from the choice of platform (i.e., MTurk) and the pre-qualification (i.e., gold question). This limits the capabilities of CI. Future research should therefore utilize multiple platforms instead of one. Instead of determining the validity of the entries by using a GQ, it is possible to select the respondents based on a timeframe for validation (e.g., set the time limit to two times the standard deviation of the mean of a trusted sample). This will serve to remove the hasty or otherwise biased answers from the dataset.

As was stated earlier, the respondents did not complete the full BWM due to limited resources, as the weights of the criteria were not established by the crowd but by experts. Hence, future studies should investigate whether the crowd could perform similarly to experts when deciding the importance of every factor influencing the battle.

Finally, human error might inadvertently occur, which may have resulted in slips and lapses. Luckily, this limitation can be easily solved in future studies by automating the calculations of the consistency ratio and the performance scores instead of manual calculations.

Analyzing and comparing CI

In this research, differences in the PERF of the crowd for the two WTT were found, leading to several recommendations for upcoming research. For instance, grading the two technologies could be done separately to determine if the same difference in PERF occurs again. If not, it would suggest that the entire task is too complex to be completed by random individuals. In addition, researchers could also investigate the influence of the monetary incentives concerning the PERF of the crowd. Therefore, when testing, future studies could follow the same approach as this research (i.e., every respondent grades both the technologies). If this approach is preferred, it is recommended to have a completion time per grading task. Doing so will allow to see whether the respondents have completed the second task faster or required a similar amount of time compared to the first grading task.

In order to solve the limitation in applying inferential statistics, it is recommended that future studies apply and compare multiple cases involving standard battles to strengthen or weaken the findings of this research and, subsequently, the theory on CI. Furthermore, more research is required to establish if the moderation effect is indeed insignificant. As was stated before, the restricting factor in this case was the limited number of DIV scores. Hence, future studies should involve more respondents to have sufficient data points, resulting in more reliable findings. In addition, a relationship between GS and DIV was established. What this relationship entails could not be tested. Therefore, future research is needed to establish what this relationship means (e.g., mediation effect).

Chapter 7

Conclusion: Reflecting this research

Chapter overview

In this chapter, the research is summarized in a general sense to give readers a final impression of the work. Discussing specific results and interpreting the data in detail is not the goal of this chapter. Instead, broad statements are elaborated that sum up the most important insights of this research. Furthermore, an empirical scientific exploratory study was conducted. Hence, the conclusion in the section below will state the main findings and recommendations of this research concisely.

- *Section 7.1 has set out to solve a practical problem with empirical research, resulting in the conclusion for the main research question. Subsequently, section 7.2 discusses the scientific contributions as well as the managerial relevance of this study.*

7.1 Conclusion

The main goal of this research is to test whether the outcome of a standard battle can be predicted by CI and how this differs from experts. In particular, the prediction that was done by the expert pool, by grading the WTTs based on pre-selected determinants. The way CI operates depends on various factors, such as survey completion time, consistency ratio, selection of best and worst criteria, PERF grade, and the final prediction. This study investigated if and how GS and DIV influenced the PERF of the collective.

Based on the five factors mentioned above, this research concluded that the crowd did show differences in the prediction of standard battles. Especially for the PERF score, as this deviated for almost two-thirds when compared to the expert pool. Despite this difference, the outcome of the prediction of the crowd was still similar to that of the experts. The consistency ratio remains a matter of doubt since most of the results were initially unreliable, making it necessary to make (logical) corrections to obtain reliable results. Consequently, this makes the comparison based on the consistency ratio rickety.

These findings may have been limited because of the differences between the RP for the GB and DD. More specifically, almost everyone performed better for the first grading task, which was the GB technology. This could indicate that the first task was too complicated, thus depleting the respondents of the resources needed to complete the second task. The monetary incentive offered when completing the survey could also have influenced the difference in PERF. For example, some respondents could have concluded after finishing the first task that the monetary incentive was too low based on the complexity of the task. Hence, depleting the motivation of the respondents. Nevertheless, to still get the reward, the respondent would speed up the second part of the survey to obtain a ‘fair’ reward—resulting in the difference in PERF. In addition, the assumption that the task was finely granulated was incorrect. Therefore, it is recommended to consider the task of predicting battles rather complex, thus harder to solve than initially thought. Subsequently, the quantity of the monetary incentive should be reconsidered and adjusted according to the complexity of the task.

The contrived groups and the individuals that form the crowd were compared based on the same factors mentioned above. The results revealed that the variance for groups is very similar compared to individuals, suggesting that the assertions by (Chiu et al., 2014; Kittur et al., 2008; Kozinets et al., 2008; Lorenz et al., 2011) are not supported because the group did not perform better than an individual.

Based on the results from the non- and parametrical test for DIV. This research clearly illustrates that DIV does not influence the PERF of the crowd. Hence, the hypothesis that a more diverse group of people should perform better was refuted. This raised the question as to whether the sample was diverse enough to grasp the capabilities of CI. In hindsight, this was limited because most users from MTurk reside in The Americas. In addition, the 'gold question,' which was included in the survey to allow pre-qualification of the respondents, was deemed too complicated or specific because the initial sample size was reduced by roughly 30%. Thus, future studies should utilize multiple platforms and exclude respondents based on a timeframe instead of a gold question.

This research also investigated the effects that GS and DIV could have on the PERF of the collective. The findings for GS were only deemed significant in the case of the GB WTT, however both the OWAT and KWT contradict each other as well as the hypothesis. Hence, the initial hypotheses were all refuted. This research was not able to conclude how these variables affect the PERF of the crowd. Nevertheless, this result weakens the theory of (S. Krause et al., 2011; Nguyen et al., 2018; Surowiecki, 2005), who underlined the importance of DIV. In contrast, the results gave ample support to (Reynolds et al., 2017) and their claim that there is no correlation between DIV and PERF. However, their second assertion about an existing relationship between cognitive DIV was not tested in this study. Hence, this was recommended for future studies.

For GS, the (non) parametrical test results indicate that a group of 10 or 15 individuals outperformed the groups with other sizes. This supports the claim of (Carvalho et al., 2016) and weakens (S. Krause et al., 2011). Another contradiction is found regarding the shape of the linear relationship, whether this was U-shaped or declining. Based on the OWAT results, the former is confirmed and thus, strengthening the claim of (Hashmi, 2005). However, for the KWT, the results contradict this claim and supports (Green, 2015), but only if the size of the group is a multiple of the OGS and not when the size increases.

Finally, using only one case limits the generalizability of the results and inhibits deeper insights. Although these limitations did not influence the findings, the results indicated no significant moderation effect, which contradicted the initial hypothesis that the relationship between GS and PERF is positively moderated by how diverse the crowd was. The findings were different to the expectations, which probably occurred since the preference was to measure DIV instead of cognitive DIV. Despite this, the results suggested that there was no relationship between DIV and PERF, thus no moderation effect was established. However, the results demonstrated that there was a significant relationship between the GS and DIV. Due to restricted resources however, this relationship was not further examined and has been recommended for future research.

7.2 Relevance of this research

The following sections discuss the contributions to the academic fields, society, and practical (managerial) implications. In addition, it is essential to acknowledge that any answer given in this chapter will be historically and contextually contingent. The same can be stated on how this research is made relevant and to what end. For instance, the relevance of scientific research may be correctly understood. Yet, how actors act upon these findings could vary differently between those producing and those consuming research. Also, rather than research being relevant upon completion, it can become relevant after the fact. This can occur in both expected and unexpected ways. Consequently, this entails that researchers have limited control over what will or will not be relevant in the future. Nevertheless, an attempt was made.

7.2.1 Scientific Relevance

This study contributed to the literature on technological forecasting and MCDM. Until now, predictions for standard battles were only based on experts' opinions from a particular industry, thus no prior research applied the concept of CI in this context. In addition, multiple case studies applied the BWM (Chitsaz et al., 2017; Gupta et al., 2016; Ren et al., 2017; Rezaei, 2015b, 2016). (Chitsaz et al., 2017; Gupta et al., 2016; Ren et al., 2017; van de Kaa et al., 2020). However, it has not been applied frequently, let alone in combination with CI. Thus, this research has demonstrated the applicability of CI in predicting technologies, and demonstrated that the BWM can be converted into an online survey and conduct a controlled study on the PERF of the crowd. Although this was an exploratory research, it still provides a new comprehensive perspective towards the current state of the art that could function as a starting point for future studies.

As for the literature on CI and CS, this research has aided in the understanding of CI and CS's disentanglement, while also providing clarification into what constitutes the phenomenon of CI. This could eventually support the theory building of CI, as there is still no consensus on which factors affect this phenomenon and to what degree. Another point that stood out when investigating other studies on CI was that very few performed an empirical study. Hence, this research contributed to the literature by providing the basis for future empirical studies. As for the application of CI, various studies have shown multiple purposes. Nevertheless, similar to the literature on MCDM, no research was found that combined these two. Hence, this research contributed by illustrating the applicability of CI in the context of MCDM and standard battles.

In the literature, many ambiguous and contradicting claims were found when optimizing the PERF of the crowd. For instance, how big the size of the contrived group should be to perform similar to experts, but also what constitutes DIV and how it influences the PERF of the crowd. This research tried to investigate which theories should be supported or challenged—using a non- and parametrical test (i.e., OWAT & KWT) for the optimal group size. Both tests indicated different but closely related OGS, for a group consisting of respectively 10 and 15 individuals. These findings suggest that the OGS lies within these values. This makes the claim of (Carvalho et al., 2016) plausible, which asserted this number to be around 10 or 11 people—subsequently, refuting the claim of 25 agents as the optimal size (S. Krause et al., 2011).

The theory about a U-shape relationship (Hashmi, 2005) between PERF and GS was tested based on the KWT and OWAT. However, both findings contradict each other. For the OWAT, the PERF improves significantly when increasing the GS of 10 to 15. But when increased from 15 to 20 members, the PERF decreases significantly—indicating a U-shape relationship. Hence, strengthening the claim of (Hashmi, 2005). In contrast, the KWT results refute (Hashmi, 2005) claim because the results indicate a linear relationship between GS and their PERF instead of the U-shape relationship. Consequently, the claim of (Green, 2015) may be valid. Even though these findings contradict each other, they also suggest there might be more than one optimum. The PERF slightly decreases for every increase of size by a multitude of the optimal size (i.e., 15). Even though this could be possible, this was not verified in this research.

As for the contribution to the theory on DIV and PERF, the OWAT and KWT results did not indicate any significant relationship. These findings simultaneously challenge and support different claims. For the former, the theory of (S. Krause et al., 2011; Nguyen et al., 2018; Surowiecki, 2005) underlined the importance of DIV in respect to the PERF. In contrast, the results give indications that the claim of (Reynolds et al., 2017) may be valid. They concluded that there was no correlation between DIV and PERF. But that there was indeed a strong correlation between cognitive DIV and PERF. This research did not test for cognitive DIV, thus this assertion is left unaffected. However, there is ample support that DIV is indeed not the magic bullet.

7.2.2 Societal & Practical Relevance

Throughout human history, elites of society used to think of the crowd as creators of problems, which later shifted to problem solvers. This is reflected by an active research area and the broad spectrum of industries leveraging the crowd. Due to globalization, decentralization, and the rapid development of new technologies, it is more than likely that this trend will continue in the future. Although this research did not investigate all the factors and merged the literature into one, it still provides value in other areas. For instance, no literature mentioned the paradox between independence and providing background information in the survey. As with every paradox, there is no clear-cut solution. Hence, more research is needed to pinpoint which kind of information researchers or institutions should avoid or include in the future when surveying predictions based on CI.

The core characteristics of the sample can be described as male, relatively young, highly educated, full-time job, and lives in the Americas. However, these characteristics limited the diversity in the sample, which was probably influenced by only using one platform (i.e., MTurk) and the boundary condition (i.e., gold question). This could have limited the full capabilities of CI. Therefore it is recommended to utilize multiple platforms instead of one. In addition, instead of the gold question, another boundary condition should be examined, for example, by selecting the respondents based on a timeframe for validation. This will allow to pinpoint and remove hasty respondents, resulting in data that is more truthful.

There were also differences in the performance of the crowd. Suggesting that the task of completing the prediction may be too complex. This assumption was validated based on the differences found when the crowd was grading the two WTT. Due to the limited number of diversity scores, more research is required to establish if the moderation effect is indeed insignificant. In addition, future individuals, studies, organizations should involve more respondents to have sufficient data points to minimize the restricting factor of DIV scores.

The only factor in which the crowd exceeded the performance of the experts was the time required to complete the survey versus the time required to conduct expert interviews. By comparing the consistency ratio, selection of the best determinant, and the final prediction drawn from the performance score, this research concluded that the crowd could perform similar to experts. However, this was not the case when selecting the worst criteria and grading the technology because the grade given by the individuals and the crowd was two-thirds lower than the grade given by experts. It can be concluded that there is support in the claim of (Larrick et al., 2012; Mollick et al., 2016; Ray, 2006; Sunstein, 2008; Tetlock, 2017) but only in certain areas.

Finally, this research also compared the results on an individual as well as group level. It was concluded that the group did not show any superior decision-making compared to the individual. This was based on the survey's duration, consistency ratio, selection of the best and worst criterion, and comparing the final performance score per alternative technology.

7.2.3 Academic Relevance

Over two years, the "Management of Technology" program teaches its graduates the different aspects appropriate for future technology managers. The focus of this program entails aspects such as personalities, analytical reasoning, organizations, assessments of technology, and how to handle human assets and technology. In combination that in all fields, continuous technological developments occur. The main qualities that this curriculum instills are devised as follows:

- **Understanding technology as a corporate resource or understood from a corporate perspective**
- **Report on scientific studies in a technological context**
- **Using scientific methods and techniques to analyze a problem as put forward in the Management of Technology curriculum**

This research investigated the prediction of standard battles for advanced technologies by utilizing the CI of a crowd that emerges by combining their individual opinions. This perspective views predictions using CI as a corporate resource and understanding from a scientific perspective.

This perspective was realized by reviewing academic literature to understand what constitutes and affects CI and how they perform relative to experts when making predictions. The contribution added to this understanding is the consequence of all-encompassing scientific research and elaborate quantifiable data analysis. The curriculum of Management of Technology contributed by cultivating these scientific skills, by courses such as "MOT2312 – Research Methods", "MOT9591 – Technology Battles", "MOT1435 – Technology, Strategy, & Entrepreneurship" and "MOT1451 – Inter- & Intra-Organizational Decision-Making".

These courses combined helped build an understanding of dominant designs, innovation, MCDM, and without research methods conducting this research would be nearly impossible. Finally, during these two years, the teachers and fellow students were a significant source of creativity and inspiration during this project.

Chapter 8

Appendix

Chapter overview

Appendix A until G was not included in this file to reduce the size of this document. However, if required to view the complete data set. Please contact me under the following address: idjordjevski@student.tudelft.nl

A: Survey - BWM Questionnaire

B: Survey – Codebook

C: Assumptions – Diversity & Relative Performance

D: Assumptions – Crowd Size & Relative Performance

E: Results Crowd Size & Performance

F: Results Crowd Diversity & Performance

G: Results Moderation Effect

Bibliography

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. doi: 10.1109/TKDE.2005.99

Allahbakhsh, M., Arbabi, S., Shirazi, M., & Motahari-Nezhad, H.-R. (2015). A Task Decomposition Framework for Surveying the Crowd Contextual Insights. *2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA)*, *October*, 155–162. doi: 10.1109/SOCA.2015.32

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing*, *17*(2), 76–81. doi: 10.1109/MIC.2013.20

Almomani, O., Almaiah, M. A., Alsaaidah, A., Smadi, S., Mohammad, A. H., & Althunibat, A. (2021). Machine Learning Classifiers for Network Intrusion Detection System: Comparative Study. *2021 International Conference on Information Technology (ICIT)*, 440–445. doi: 10.1109/ICIT52682.2021.9491770

Amazon. (2021). *MTurk*. Retrieved from https://www.mturk.com/help

Anantapantula, A. (2017). *Factors affecting Standard Dominance in the battle between EDIFACT versus XBRL Data Exchange Standards in India*.

Aris, H. (2017). Current State of Crowdsourcing Taxonomy Research: a Systematic Review. *ICOCI Kuala Lumpur. Universiti Utara Malaysia*, *176*, 25–27.

Aris, H., & Din, M. M. (2016). Crowdsourcing evolution: Towards a taxonomy of crowdsourcing initiatives. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 1–6. doi: 10.1109/PERCOMW.2016.7457122

Benkler, Y. (2006). The Wealth of Networks: How Social Production Transforms Markets and Freedom. *Social Science Computer Review*. doi: 10.1177/1084713807301373

Bhatti, S. S., Gao, X., & Chen, G. (2020). General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey. *Journal of Systems and Software*, *167*, 110611. doi: 10.1016/j.jss.2020.110611

Bonabeau, E. (2009). Decisions 2.0 - the Power of Collective Intelligence. *Mit Sloan Management Review*. doi: DOI 10.1007/s12599-010-0114-8

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 75–90. doi: 10.1177/1354856507084420

Budescu, D. V., & Chen, E. (2015). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*, *61*(2), 267–280. doi: 10.1287/mnsc.2014.1909

Carvalho, A., Dimitrov, S., & Larson, K. (2016). How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, *78*(1), 45–72. doi: 10.1007/s10472-015-9492-4

Chen, J., Ren, Y., & Riedl, J. (2010). The effects of diversity on group productivity and member withdrawal in online volunteer groups. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, *2*, 821. doi: 10.1145/1753326.1753447

Chesbrough, H. (2003). *Open Innovation*. doi: 10.1016/j.respol.2004.08.005

Chitsaz, N., & Azarnivand, A. (2017). Water Scarcity Management in Arid Regions Based on an Extended Multiple Criteria Technique. *Water Resources Management*, *31*(1), 233–250. doi: 10.1007/s11269-016-1521-5

Chiu, C.-M., Liang, T.-P., & Turban, E. (2014). What can crowdsourcing do for decision support? *Decision Support Systems*, *65*(C), 40–49. doi: 10.1016/j.dss.2014.05.010

Corney, J. R., Torres-Sánchez, C., Jagadeesan, A. P., Yan, X. T., Regli, W. C., & Medellin, H. (2010). Putting the crowd to work in a knowledge-based factory. *Advanced Engineering Informatics*, *24*(3), 243–250. doi: 10.1016/j.aei.2010.05.011

Cramer, D. (2003). *Fundamental Statistics for Social Research*. Routledge. doi: 10.4324/9780203360613

Cramer, D., & Howitt, D. (2004). *The SAGE Dictionary of Statistics*. 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications, Ltd. doi: 10.4135/9780857020123

Doane, D. P., & Seward, L. E. (2011). Measuring Skewness: A Forgotten Statistic? *Journal of Statistics Education*, *19*(2). doi: 10.1080/10691898.2011.11889611

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. doi: 10.1177/0165551512437638

Felton, L., & Jowett, S. (2013). Attachment and well-being: The mediating effects of psychological needs satisfaction within the coach-athlete and parent-athlete relational contexts. *Psychology of Sport and Exercise*, *14*(1), 57–65. doi: 10.1016/j.psychsport.2012.07.006

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gao, J., Li, Q., Zhao, B., Fan, W., & Han, J. (2015). Truth discovery and crowdsourcing aggregation. *Proceedings of the VLDB Endowment*, *8*(12), 2048–2049. doi: 10.14778/2824032.2824136

Glass, G. V, Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, *42*(3), 237–288. doi: 10.3102/00346543042003237

Green, B. (2015). *How can we measure Collective Intelligence?* Unanimous AI Blog. Retrieved from https://unanimous.ai/can-measure-collective-intelligence/

Gupta, H., & Barua, M. K. (2016). Identifying enablers of technological innovation for Indian MSMEs using best–worst multi criteria decision making method. *Technological Forecasting and Social Change*, *107*, 69–79. doi: 10.1016/j.techfore.2016.03.028

Gurari, D., Sameki, M., & Betke, M. (2016). Investigating the Influence of Data Familiarity to Improve the Design of a Crowdsourcing Image Annotation System. *In 4th AAAI Conf. Human Comput. and Crowdsourc. (HCOMP)*, 59–68.

Gyulavári, T. (2020). Collective rights of platform workers: The role of EU law. *Maastricht Journal of European and Comparative Law*, *27*(4), 406–424. doi: 10.1177/1023263X20932070

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *Journal of Educational Statistics*, *17*(4), 315–339. doi: 10.3102/10769986017004315

Hashmi, N. (2005). *The more the merrier? : understanding the effect of group size on collective intelligence*. 104. Retrieved from https://dspace.mit.edu/handle/1721.1/113957

Hong, L., & Page, S. E. (2012). Some Microfoundations of Collective Wisdom. In H. Landemore & J. Elster (Eds.), Collective Wisdom (pp. 56–71). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511846427.004

Howe, J. (2006). The Rise of Crowdsourcing. *Wired*. Retrieved from https://www.wired.com/2006/06/crowds/

Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia. *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work - CSCW '08*, *3*(0), 37. doi: 10.1145/1460563.1460572

Kozinets, R. V., Hemetsberger, A., & Schau, H. J. (2008). The Wisdom of Consumer Crowds. *Journal of Macromarketing*, *28*(4), 339–354. doi: 10.1177/0276146708325382

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, *25*(1), 28–34. doi: 10.1016/j.tree.2009.06.016

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948. doi: 10.1016/j.anbehav.2010.12.018

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The Social Psychology of the Wisdom of Crowds. *Natural Language Engineering*.

Levy, P., & Bononno, R. (1997). *Collective INtelligence: Mankind's Emerging World In Cyberspace*. Perseus Books.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of Assumption Violations Revisited:

A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research*, *66*(4), 579–619. doi: 10.3102/00346543066004579

Lorenz, J. (2021). *On the Quantification of Crowd Wisdom*. doi: 10.31234/osf.io/6ydg4

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. doi: 10.1073/pnas.1008636108

Malone, T.W., & Bernstein, M. S. (2015). *Handbook of Collective Intelligence*. MIT Press. Retrieved from https://mitpress.mit.edu/books/handbook-collective-intelligence

Malone, T.W., & Levy, P. (2008). *Collective Intelligence: Creating a Prosperous World at Peace*. Earth Intelligence Network.

Malone, Thomas W., Laubacher, R., & Dellarocas, C. N. (2009). Harnessing Crowds: Mapping the Genome of Collective Intelligence. *SSRN Electronic Journal*. doi: 10.2139/ssrn.1381502

Malone, Thomas W, Work, O., & Like, L. (2003). The Future of Work Research Programme. *Work, Employment & Society*, *17*(3), 553–556. doi: 10.1177/09500170030173008

Matzler, K., Strobl, A., & Bailom, F. (2016). Leadership and the wisdom of crowds: how to tap into the collective intelligence of an organization. *Strategy & Leadership*, *44*(1), 30–35. doi: 10.1108/SL-06-2015-0049

McGrath, J. E. (1984). Groups: Interaction and Performance. In Prentice-Hall.

Milgram, S. (1963). Behavioral Study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. doi: 10.1037/h0040525

Mohd Razali, N., & Bee Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 13–14.

Mollick, E., & Nanda, R. (2016). Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. *Management Science*, *62*(6), 1533–1553. doi: 10.1287/mnsc.2015.2207

Nakatsu, R. T., Grossman, E. B., & Iacovou, C. L. (2014). A taxonomy of crowdsourcing based on task complexity. *Journal of Information Science*, *40*(6), 823–834. doi: 10.1177/0165551514550140

Nalmpantis, D., Roukouni, A., Genitsaris, E., Stamelou, A., & Naniopoulos, A. (2019). Evaluation of innovative ideas for Public Transport proposed by citizens using Multi-Criteria Decision Analysis (MCDA). *European Transport Research Review*, *11*(1), 22. doi: 10.1186/s12544-019-0356-6

Nguyen, V. Du, & Nguyen, N. T. (2018). Intelligent Collectives: Theory, Applications, and Research Challenges. *Cybernetics and Systems*, *49*(5–6), 261–279. doi: 10.1080/01969722.2017.1418254

Oliver, P. E., & Marwell, G. (1988). The Paradox of Group Size in Collective Action: A Theory of the Critical Mass. II. *American Sociological Review*, *53*(1), 1. doi: 10.2307/2095728

Page, S. E. (2007). *The Difference*. Princeton University Press.

Ram, B. Y. (2018). *Factors for Dominance during a Standards Battle general*. TUDelft.

Ray, R. (2006). Prediction Markets and the Financial "Wisdom of Crowds." *Journal of Behavioral Finance*, *7*(1), 2–4. doi: 10.1207/s15427579jpfm0701_1

Ren, J., Liang, H., & Chan, F. T. S. (2017). Urban sewage sludge, sustainability, and transition for Eco-City: Multi-criteria sustainability assessment of technologies based on best-worst method. *Technological Forecasting and Social Change*, *116*, 29–39. doi: 10.1016/j.techfore.2016.10.070

Reynolds, A., & Lewis, D. (2017). Solve Problems Faster When They're More Cognitively Diverse. *Harvard Buisness Review*, *March*, 2–7.

Rezaei, J. (2015a). Best-worst multi-criteria decision-making method. *Omega (United Kingdom)*, *53*, 49–57. doi: 10.1016/j.omega.2014.11.009

Rezaei, J. (2015b). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57. doi: 10.1016/j.omega.2014.11.009

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, *64*, 126–130. doi: 10.1016/j.omega.2015.12.001

Robert, L. P., & Romero, D. M. (2017). The influence of diversity and experience on the effects of crowd size. *Journal of the Association for Information Science and Technology*, *68*(2), 321–332. doi: 10.1002/asi.23653

Robert, L., & Romero, D. M. (2015). Crowd Size, Diversity and Performance. *Proceedings of the 33rd*

Annual ACM Conference on Human Factors in Computing Systems, *2015-April*, 1379–1382. doi: 10.1145/2702123.2702469

Rosen, P. A. (2011). Crowdsourcing Lessons for Organizations. *Journal of Decision Systems*, *20*(3), 309–324. doi: 10.3166/jds.20.309-324

Salminen, J. (2015). *The Role of Collective Intelligence in Crowdsourcing Innovations*.

Schenk, E., & Guittard, C. (2009). Crowdsourcing : What can be Outsourced to the Crowd , and Why ? *Innovation*, *January*, 1–29.

Schulze, T., Seedorf, S., Geiger, D., Kaufmann, N., & Schader, M. (2011). Exploring task properties in crowdsourcing - An empirical study on Mechanical Turk. *19th European Conference on Information Systems, ECIS 2011*, *January 2011*.

SHAPIRO, S. S., & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. doi: 10.1093/biomet/52.3-4.591

SIMPSON, E. H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688–688. doi: 10.1038/163688a0

Skaržauskiene, A., & Mačiuliene, M. (2015). Modelling the index of collective intelligence in online community projects. *Proceedings of the 10th International Conference on Cyber Warfare and Security, ICCWS 2015*, 313–319.

Sloane, P. (2011). *A Guide to Open Innovation and Crowdsourcing*. Kogan Page.

Solver Linear BWM. (2016). Retrieved from https://bestworstmethod.com/software/

Sunstein, C. R. (2008). *Infotopia: How Many Minds Produce Knowldege*.

Suran, S., Pattanaik, V., & Draheim, D. (2020). Frameworks for Collective Intelligence. *ACM Computing Surveys*, *53*(1), 1–36. doi: 10.1145/3368986

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the MAny are Smarten than the Few and How Collective Wisdom Shapes Business*.

Tapscott, D., & Williams, A. D. (2008). *Changes Everything*. doi: https://doi.org/10.1111/j.1460-2466.2008.00391_5.x

Tedjakusuma, V. (2014). *THE EMERGENCE OF DE FACTO STANDARDS: The value of integrative frameworks in the analysis of standards battles*. TUDelft.

Tehrani, M. Q. (2014). *Negative Indirect Network Effects*. TUDelft.

Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know?*

Tyler, T. R., & Lind, E. A. (1992). A relational model of authority in groups. *Advances in Experimental Social Psychology*, *25*(C), 115–191. doi: 10.1016/S0065-2601(08)60283-X

v.d. Kaa, G., v.d. Ende, J., de Vries, H.J., van Heck, E. (2011). Factors for winning interface format battles: A review and synthesis of the literature. *Technological Forecasting and Social Change*, *78*(8), 1397–1411. doi: 10.1016/j.techfore.2011.03.011

van de Kaa, G., Fens, T., & Rezaei, J. (2019). Residential grid storage technology battles: a multi-criteria analysis using BWM. *Technology Analysis & Strategic Management*, *31*(1), 40–52. doi: 10.1080/09537325.2018.1484441

van de Kaa, G., Kamp, L., & Rezaei, J. (2017). Selection of biomass thermochemical conversion technology in the Netherlands: A best worst method approach. *Journal of Cleaner Production*, *166*, 32–39. doi: 10.1016/j.jclepro.2017.07.052

van de Kaa, G., van Ek, M., Kamp, L. M., & Rezaei, J. (2020). Wind turbine technology battles: Gearbox versus direct drive - opening up the black box of technology characteristics. *Technological Forecasting and Social Change*, *153*, 119933. doi: 10.1016/j.techfore.2020.119933

Vander Schee, B. A. (2009). Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business20093Jeff Howe. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business . New York, NY: Crown Business 2008. 320 pp. $26.95. *Journal of Consumer Marketing*, *26*(4), 305–306. doi: 10.1108/07363760910965918

von Hippel, E. (1986). Lead Users: A Source of Novel Product Concepts. *Management Science*, *32*(7), 791–805. doi: 10.1287/mnsc.32.7.791

Wagner, C., & Vinaimont, T. (2010). Evaluating the wisdom of crowds. *Proceedings of Issues in Information Systems*, *XI*(1), 724–732. Retrieved from https://pdfs.semanticscholar.org/b6d9/4936607ff08308b794ae60c26e9e8f5c42eb.pdf

Wexler, M. N. (2011). Reconfiguring the sociology of the crowd: Exploring crowdsourcing. *International Journal of Sociology and Social Policy*, *31*(1–2), 6–20. doi: 10.1108/01443331111104779

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective Intelligence and Group Performance. *Current Directions in Psychological Science*, *24*(6), 420–424. doi: 10.1177/0963721415599543

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, *330*(6004), 686–688. doi: 10.1126/science.1193147

Yu, C., Chai, Y., & Liu, Y. (2017). Collective Intelligence. *Proceedings of the 2nd International Conference on Crowd Science and Engineering - ICCSE'17*, *Part F1306*, 111–115. doi: 10.1145/3126973.3126993

Zaarour, B. (2011). *The Interrelation of Factors for Standard Dominance in Standard Battles*. TUDelft.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, *16*(3), 417–434. doi: 10.1007/s10796-012-9350-4