



School of Technologies and Architecture, ISCTE  
Faculty of Sciences, University of Lisbon

# An experiment about the impact of social influence on the wisdom of the crowds' effect

Sofia Silva  
(asgss@iscte.pt)

Dissertation submitted in partial requirement for the  
conferral of Master of Sciences degree in Complexity Sciences

Supervisor:  
Professor Luís Correia  
Faculty of Sciences, University of Lisbon

June 2016

# Acknowledgments

I would like to express my profound gratitude to my dearest friend István Mandak, without whom this dissertation would have not been possible. He contributed with his knowledge and ideas to implement the code for the experiment and committed his time to improve the way data was collected. Preventing duplicates in the framework of MTurk has turned out to be exceptionally challenging and his long term support and patience was invaluable to achieve a reliable setup. I equally wish to thank his wife, Adrienn and his two kids, Maja and Bohus, who inadvertently accompanied this project and with whom I ended sharing dinner, games and a great amount of jellybeans.

I would like to sincerely thank my supervisor Luís Correia who patiently guided me throughout this bumpy and rather long journey. I believe that without his continuous support, patience, and readiness to communicate (even if three thousand kilometres away) it would have been extremely difficult to overcome the many challenges of this project. His continuous enthusiasm and good humour were truly motivating.

I also want to thank Jorge Louçã for his initial encouragement and support when I first started the program. Furthermore, I would like to thank my dearest friends Sabrina Amendoeira, Diana Ferreira, and Juergen Krasser who have always been encouraging and supportive.

*To my father*

# Table of Contents

<b>1. Introduction.....</b>	<b>1</b>
1.1.1. <i>Motivation and goal.....</i>	1
1.1.2. <i>Hypothesis.....</i>	2
1.1.3. <i>Outline of the dissertation.....</i>	3
<b>2. Literature review .....</b>	<b>4</b>
2.1. Conceptual foundations .....	4
2.1.1. <i>First evidences of the wisdom of the crowds.....</i>	5
2.1.2. <i>Condorcet's Jury Theorem.....</i>	5
2.1.3. <i>A working definition of the wisdom of the crowds.....</i>	6
2.1.4. <i>Collective behaviour is not necessarily intelligent.....</i>	6
2.2. Diversity .....	7
2.2.1. <i>What is diversity?.....</i>	7
2.2.2. <i>Deconstructing Cognitive Diversity.....</i>	10
2.2.3. <i>Diversity trump ability theorem.....</i>	11
2.3. Measures .....	13
2.3.1. <i>Aggregation methods.....</i>	13
2.3.2. <i>Considerations about aggregation.....</i>	15
2.3.3. <i>Assessing the strength of WoC effect: Wisdom of the crowd indicator.....</i>	15
2.4. Typology of problems .....	16
2.4.1. <i>Estimation of events.....</i>	18
2.4.2. <i>Problem solving: single variable.....</i>	19
2.4.3. <i>Decision making: multivariable.....</i>	19
2.5. Social influence .....	23
2.5.1. <i>Context.....</i>	23
2.5.2. <i>Social influence impact on the wisdom of the crowd effect.....</i>	24
2.5.3. <i>Social interaction as an aggregation mechanism.....</i>	25
2.5.4. <i>The impact of imitation.....</i>	27
<b>3. Methods .....</b>	<b>29</b>
3.1. Experiment Design.....	29
3.1.1. <i>Introduction.....</i>	29
3.1.2. <i>Set up of the experiment.....</i>	30
3.2. Collection and treatment of data.....	34
3.2.1. <i>Data collection platform: Amazon Mechanical Turk.....</i>	34
3.2.2. <i>Reliability of data.....</i>	35
3.2.3. <i>Erased files.....</i>	35
3.2.4. <i>Considerations about erased files.....</i>	36
<b>4. Results .....</b>	<b>37</b>
4.1. Individual group analysis .....	37
4.1.1. <i>Group 0: no influence.....</i>	37
4.1.2. <i>Group 1: the 5 best out of 10 random estimates.....</i>	38
4.1.3. <i>Group 2: bracketed mode.....</i>	40
4.1.4. <i>Group 3: full information.....</i>	41
4.2. Group performance analysis.....	43
4.2.1. <i>Arithmetic mean.....</i>	43
4.2.2. <i>Measures from Diversity Prediction Theorem: Individual error.....</i>	43
4.2.3. <i>Measures from Diversity Prediction Theorem: Diversity and Collective Error.....</i>	44
4.2.4. <i>Dispersion of estimates: Wisdom of the Crowd Indicator.....</i>	45

<b>5. Conclusions .....</b>	<b>47</b>
5.1. Theoretical implications.....	47
5.2. Future research.....	49
<b>6. References.....</b>	<b>50</b>
<b>7. Appendices.....</b>	<b>52</b>
Appendix A: List of erased values.....	52
Appendix B: Levene's test.....	53
Appendix C: Shapiro-Wilk test for normality .....	54
Appendix D: Mann-Whitney U-Tests.....	55
Appendix E: Wisdom of the crowd indicator .....	58
Appendix F: Code PHP/HTML of experiment page.....	59
Appendix G: Experiment layout of group 0 .....	60
Appendix H: Experiment layout of group 2 .....	61
Appendix I: Experiment layout of group 2.....	62
Appendix J: Experiment layout of group 3 .....	63

# List of figures, tables and equations

Figure 1: Heuristic box .....	12
Figure 2: McGrath's circumplex model.....	16
Figure 3: Conformity experiment.....	23
Figure 4: information degrees in groups.....	29
Figure 5: Experiment landing page.....	30
Figure 6: Survey screen.....	31
Figure 7: Experiment flow .....	33
Figure 8: Distribution and global statistics for group 0.....	37
Figure 9: Distribution of estimates for group 0 .....	38
Figure 10: Distribution and global statistics for group 1 .....	39
Figure 11: Distribution of estimates for group 1 .....	39
Figure 12: Distribution and global statistics for group 2.....	40
Figure 13: Distribution of estimates for group 2 .....	40
Figure 14: Global Statistics and distribution for group 3.....	41
Figure 15: Distribution of estimates for group 3 .....	41
Figure 16: Group distribution overview.....	43
Figure 17: Individual error over time.....	44
Table 1: Average of predictions .....	8
Table 2: Expression of diversity .....	8
Table 3: Diversity Prediction Theorem .....	9
Table 4: McGrath's Quadrants.....	16
Table 5: Comprehensive WoC literature review .....	21
Table 6: List of values orderered by input in group 0 .....	38
Table 7: List of values orderered by input in group 1 .....	39
Table 8: List of values orderered by input in group 2 .....	40
Table 9: List of values orderered by input in group 3 .....	42
Table 10: Diversity in all groups .....	45
Table 11: Wisdom of the Crowd Indicator .....	58
Equation 1 - Diversity Prediction Theorem equation .....	9

# Abstract

Groups have the impressive ability to perform better collectively than the best of its individuals. Galton observed this first in 1907 in his ox weight experiment, but the term *wisdom of the crowds* (WoC) was coined only later in 2004 by Surowiecki. Cognitive diversity at the individual level enables groups to produce differentiated solutions that ultimately cluster near the true value. By cancelling out the wrongs, the aggregation method exposes the convergence of multiple local optima solutions into one, typically an averaged value that comes incredibly close to the truth-value of what is being estimating.

Some accounts suggest that social influence hinders the WoC effect because it diminishes the group diversity resulting in biased outcomes. However, social influence is a naturally occurring phenomenon and it is hardly determinable the extent to which individuals are biased or independent given the complexity of the social interactions.

We investigated the impact of social influence on the WoC effect by comparing the collective predictions of 4 groups regarding the number of jellybeans in a jar. We demonstrate that the group disclosing full information performs nearly as well as the control group, where no information was shared. The aggregation method to converge the estimates was the arithmetic mean showing that both groups predicted by approximately 7% the correct number. Statistical analysis has shown that diversity is not affected significantly in the social groups.

We conclude that the WoC is not affected by social influence but by the degree of aggregation of the social information shared.

# Resumo

Um grupo de pessoas tem a impressionante capacidade de obter melhores resultados a resolver problemas como colectivo do que o mais capaz dos seus indivíduos. Galton observou este fenómeno pela primeira vez na experiência que levou a cabo em 1907 num concurso sobre o peso de um boi, embora o termo *wisdom of the crowds* (WoC) só viesse a ser popularizado mais tarde, em 2004, por Surowiecki.

A diversidade cognitiva a nível individual possibilita a criação de uma variedade de soluções ao nível colectivo que acaba por gravitar em torno do valor real uma vez que os valores errados se cancelam mutuamente quando é aplicado um método agregador, normalmente a média.

Alguns autores sugerem que a influência social dificulta o efeito de WoC porque diminui a diversidade dos grupos e por conseguinte produz resultados tendenciosos.

No entanto a influência é um fenómeno que ocorre naturalmente e é difícil determinar o grau de influência individual devido à complexidade de interações sociais.

Investigámos o impacto da influência social sobre o efeito de WoC comparando as estimativas colectivas de 4 grupos relativamente ao número de doces num jarro. Demonstrámos que o grupo que mostra informação colectiva total obtém resultados semelhantes ao grupo de controlo onde nenhuma informação é partilhada.

Usando a média aritmética, os dois grupos previram com uma eficácia aproximada de 7% o número correto de doces no jarro. Testes estatísticos revelaram que a diversidade nos grupos sob influência social não foi significativamente diferente da do grupo de controlo.

Concluimos que a influência social não interfere com a diversidade dos grupos se se manifestar de forma integral incluindo toda a informação das estimativas anteriores sem haver convergência de valores.

Keywords: influência social, wisdom of the crowds, diversidade cognitiva, inteligência colectiva.



## Abbreviation list

AE	Average individual Error
CE	Collective Error
CI	Collective Intelligence
D	Diversity
WoC	Wisdom of the Crowds

# 1. Introduction

## 1.1.1. Motivation and goal

Although the Wisdom of the Crowds - a form of collective intelligence - is not an entirely new concept, the way new technologies have enabled the access to data analysis is fairly recent. This phenomenon has been first observed by Francis Galton in 1907 in his ox weight experiment [1], but the term *wisdom of the crowds* was coined by Surowiecki in 2004 referring to the ability of crowds to solve problems or make predictions better than the best of its individuals.

Many complex real-world problems can benefit from this surge to enable shared, cooperative production of knowledge and support decision-making processes that otherwise would become too difficult to collect or to coordinate. The renewed interest in collective intelligence is owed to the dramatic new forms that communication technologies have imprinted to human social life. Currently, millions of people are connected through the Internet and there is increased mobility with the rise of smartphones, which is transforming connectedness and social habits at an incredible pace. It is possible to observe that, in only two decades, human communication has shifted from a local to a global paradigm, allowing a whole new kind of interaction possibilities which have never been possible before. The role of distributed problem solving powered by decentralized, digital platforms is therefore at the core of a new chapter for the study of collective intelligence phenomena.

As societies experience new forms of communication and become more interactive, inherently connected and increasingly complex, the need for systems to support collective awareness increases. Collective intelligent instruments, such as the wisdom of the crowds (WoC), emerge as a possible way to tackle forecasting and decision making in a growing complexity world.

One powerful motivation to investigate collective intelligence, and more concretely the Wisdom of the Crowds effect, is the immense potential and applicability to solve real-world problems, in particular the degree to which collective cognition generates better, more accurate results than groups of experts, which becomes increasingly relevant in the current technological context.

The study of the Wisdom of the Crowds (WoC) phenomenon is still at its infancy and still lacking a solid framework both at application and assessment level. The conditions that can support successful results, such as: size group, cognitive group diversity, social influence, access to information and aggregation methods, still require further research to reach an applicability standard. The literature reviewed at the present moment, expresses that ambiguity: different authors refer to the same metrics and concepts differently and the measurements for success are inconsistent throughout experiments.

The broader goal of this research work is, optimistically, to shed light onto the enabling conditions of this imminent collective intelligence tool - the wisdom of the crowds. More concretely this thesis will focus on one particular enabling condition of the wisdom of the crowds generally described as social influence.

Social influence - a very well-known phenomenon in psychology and behavioural sciences - refers to the degree to which each individual can be biased and deviated from her own original behaviour to blend with those of others in a group. Consequently, social influence plays a role on the diversity of groups: by affecting individual solutions, groups become more homogeneous and eventually lean towards a biased converged sub optimal solution.

The current view on the role of the social influence [2][3][4] is that it has a negative impact on the WoC effect because it diminishes the group diversity by producing biased solutions. Many experiments have isolated individuals to prevent social contact and for many types of problems, social isolation produces successful collective solutions. However, the limits of social influence are unclear. The extent to which individuals are biased or independent is hardly definable given the complexity of the social interactions that take place in real-world scenarios. Departing from the imprecise premise that social influence undermines the wisdom of the crowds' effect, we set ourselves to investigate the impact of social influence in the wisdom of the crowd's effect.

Two main reasons motivated us to challenge the consensual assumption that social influence undermines the WoC effect: first the notion that consensual decision in animal groups contributes to cohesion, speed and accuracy of decision-making [5]. Animal groups reach consensus by locally interacting with each other. In this exchange, the likelihood of individuals to choose one option increases with the number of others already committed to that option. The positive feedback ultimately directs the group of individuals towards the best available choice. With no central control, self-organisation explains how imitation has a positive role in the creation of heterogeneous social patterns in uniform environments; for example ants choose the shortest route to food using pheromones as a positive feedback mechanism that will recruit other ants to follow the same route. Colonies of ants and honeybees can also collectively choose the best source for food recruiting effectively according to source quality [5]. Thus, imitation does seem to play a role in achieving optimized decision making results in the context of animal behaviour.

Secondly, the experiments we reviewed refer to the impact of social influence in a very limited, unnatural setting which can hardly correspond to real life scenarios. With this in mind we decided to investigate if social influence, as positive feedback, can itself be a building block of the WoC effect.

### 1.1.2. Hypothesis

The research work outlined in this thesis aims to answer the following question: is social influence undermining the WoC effect or can it be a constructor of WoC?

More concretely we focus on different degrees of social influence to understand if non-aggregated social influence produces better results than the aggregated, hypothesizing that non-aggregated information will lead to better results by feeding more, untreated information that subjects can then willingly take into account. This means that instead of providing subjects with an externally processed aggregated value, we change the focus on to each individual's ability to aggregate based on the full scope of estimates.

Social influence has been largely suggested to have a negative impact in the WoC [2][3][4]. While social influence diminishes the diversity of groups by introducing pressure towards conformity - hence reducing the possibility for different perspectives on the same problem - non-socially influenced behaviour is extremely rare and difficult to quantify in natural environments. In fact, many social and biological systems rely on the observation of others and of the environment to adapt and revise their behaviour as a process of learning and introduction to innovation. Departing from this idea, this dissertation explores the hypothesis that social influence in its full information form - hence with no external treatment or aggregation - might have a positive contribution to the wisdom of the crowd's effect.

To deconstruct the mechanics of social influence on the WoC phenomena, we centred our experiment on a problem-solving type of challenge, with similar complexity to Galton's experiment [1] but added three groups with access to different degrees of information to test how social influence affects the performance of the groups. We included a control group with no access to information and tested against the information groups using the arithmetic mean, measures of the Diversity Prediction Theorem and statistical tests to assess the homogeneity of variance and the significance of differences in distribution.

Our hypothesis is that social influence does not undermine the WoC effect but enables it when disclosing full non-aggregated information. If groups accessing collective non-aggregated information perform as well or better than the control group - where no information is shared - then social influence does not affect the accuracy of collective performance but impacts it positively.

### 1.1.3. Outline of the dissertation

This thesis is outlined in 5 chapters. After this first chapter, we present in the second chapter the reviewed literature where we consider the essential aspects related to the WoC: diversity, measures, social influence and practical applications. Next, we describe the details of the experiment used to test our hypothesis. We explain the methods and describe the platform, handling of data and the mechanics and design of the experiment. In the fourth chapter we describe the results, first a detailed analysis for every group, then a comparison of the performance of groups followed by a discussion. Lastly, the fifth chapter concludes the research we undertook and leaves notes for future research.

## 2. Literature review

### 2.1. Conceptual foundations

*Googling* something has become perhaps one of the most common tasks in our daily lives, yet one the most elegant forms of collective intelligence. Besides showing in a fraction of seconds the most relevant results, based on PageRank<sup>1</sup> algorithm,

Google has also the capacity to correct expressions or mistakes. It *understands* meaning because it relies on a set of algorithms that reinforce *meaning* according to frequency and the strength of links between pages [6]. Researchers are learning more about how the simultaneous acquisition and information processing from distributed sources can originate high-order collective capabilities. Recently, the interest for this kind of phenomena has experienced a resurgence. The increasing connectedness of our societies and the easy access to large collections of data available from online sources has enabled new experiments to explore naturally occurring collective intelligence phenomena, such as idea spreading, coalition forming or group evolution [7]. Within the definition [8] of collective intelligence as *groups of individuals doing things collectively that seem intelligent*, the notion that collective intelligence exists for a long time is evident in organizations such as families, companies, countries and other groups, in which it is possible to grasp some sort of actions that *seem* intelligent [6].

This chapter will introduce one specific form of collective intelligence, usually referred to as *the wisdom of crowds*. This terminology has been also referred to by other authors [9] as “swarm intelligence”, “crowdsourcing”, “peer production”, “user communities”, “collective wisdom”, “distributed problem solving” or more generally as “collective intelligence” [10]. To preserve clarity, this dissertation will make use of collective intelligence (CI) as a general capacity of groups to solve cognitive problems that go beyond individual capacities [11]. The wisdom of the crowds (WoC) will be held as a specific tangible form of CI that refers to a statistical phenomenon where the collective averaged performance is more accurate than the best of its individuals, and will also be interchangeably mentioned throughout the document as *the wise effect*, *wisdom*, *intelligence*, *crowd wisdom* or *crowd intelligence*.

---

<sup>1</sup> PageRank is a link analysis algorithm created by Larry Page that measures the importance of web pages by counting the number and quality of links to a page [37].

### 2.1.1. First evidences of the wisdom of the crowds

In 1907, at a livestock fair in Plymouth, Francis Galton [1] studied the results of a weight-juggling competition. About 800 people bought stamped and numbered cards in which they inscribed the estimates of an ox weight after it would have been slaughtered and dressed. For the most successful guesses a prize would be at stake. The six penny fee prevented anecdotal bets and the pursue of the prize instilled each player to do his or her best. What Galton discovered, after analysing the submitted tickets, was that the average of all the estimates was more accurate than the winner's estimate. *Vox Populi* was then correct to within 1 per cent of the real value. The remarkable fact is that most of the betters had little or no experience at all about cattle weighting, yet the accuracy of their opinions together surpassed the accuracy of the expert opinion [12]. Galton poses an interesting analogy in his article [1] on how the average competitor is as capable of guessing the weight of an ox as an average voter is of judging the merits of political issues she votes for - *many heads are better than one* was an idea seriously defended by Aristotle in his arguments for democracy.

### 2.1.2. Condorcet's Jury Theorem

The basic insight for collective wisdom was first formalized in the 18th century by the Marquis of Condorcet, a French mathematician and political philosopher whose ideas embodied the ideals of the Age of Enlightenment and Rationalism. The Condorcet's jury theorem proved that if the number of votes is large enough, a simple 'democratic' majority vote, made independently by individuals, provided accurate results [12]. The theorem considers a situation where a jury of  $n$  members decides between two options. Each individual has a probability  $p$  of making the best decision. Given that  $p > 0.5$ , the chance of a correct collective decision of the group increases as a function of group size. In other words, if everyone meets the rather low standard of exceeding a chance probability of being correct, the group as a whole can come near to a 100% chance of making the right choice [5][10][11]. However, to achieve collective accuracy, the Condorcet's theorem requires that individuals are unbiased and independent, which can present itself as a challenge, given that decision-making processes not confined to the isolation of a ballot box do rely on communication to achieve consensual decision. So how can collective decisions preserve independence but still come to a final consensus? [5] Moreover, the theorem also requires that the majority of individuals possess the correct information so it has more to do with the means by which consensus decisions are made when a majority already has correct information [11]. But, even if  $p$  is close to one-half, when multiplied by a big crowd the difference between correct and incorrect answers increases. This phenomenon is also known as the Law of Large Numbers: the true value of  $p$  reveals itself as more independent signals get

produced; so if  $p$  is greater than half a crowd will eventually get the right answer [13].

### 2.1.3. A working definition of the wisdom of the crowds

The expression “Wisdom of the crowds” was popularized by Surowiecki’s homonymous book in 2004, in which he explores how the concepts and ideas of collective wisdom can be applied to shape businesses, society and nations. Since then, the term has been used interchangeably to refer to a number of different phenomena, commonly suggesting crowdsourcing mechanisms that do not actually involve the WoC effect itself.

Due to the fact that the WoC effect is still a very recent and an unexplored phenomenon, it lacks an accurate definition and many of its traits remain unclear or are inconsistent throughout literature. For the purpose of this dissertation, we will refer to the wisdom of the crowd effect as a specific form of swarm intelligence of statistical nature [14][2][3][15], where the averaged outcome of a group is significantly better than the best of its individual performance.

The *crowd beats experts* summarises the most distinct defining trait of the wisdom of crowds. A common example to illustrate this concept is the popular TV show *Who Wants to be a Millionaire*, where the guesses of the audience lead to the correct answer 91% of the cases [16]. In another case, the aggregation of experts’ estimations from several disciplines had been more accurate in the localization of a missing submarine than individual expert estimation [17]. The crowd also performs better in combinatorial optimization problems where participants were asked to provide their solution for optimisation problems, such as the Spanning Tree Problem and Traveling Salesman Problem [14]. The performance of the aggregated solutions is drastically better than the individual solutions, being only outperformed by approximately 2% of the participants [14].

### 2.1.4. Collective behaviour is not necessarily intelligent

A common misunderstanding regarding the intelligence of crowds comes from examples like crowd panics, riots or herd behaviour where the masses behave collectively in a non-rational manner. Although the definitions of collective behaviour and WoC suffer from a lack of clear criteria, not all types of collective behaviour can be regarded as an evidence of *intelligence*. While collective intelligence and collective behaviour exist in relation to each other, a flock of birds or individuals with socially determined motion patterns are not, in essence, examples of collective intelligence. It is evident that collective behaviour also entails some individual decision-making processes that also reach for consensual decision-making, yet this only confirms that individual decision takes into consideration the social context and that sometimes those decisions are consensual, it does not however reveal a particular form of swarm intelligence [11].

## 2.2. Diversity

### 2.2.1. What is diversity?

Diversity is a fundamental condition for the crowd to be wise. For a successful implementation of the WoC, the crowd must promote individual, differentiated opinions so that each individual's information, even if constrained to a personal interpretative notion, is diverse.

Diversity can be shaped by promoting the composition of the group (who is in the group) or by the process of sharing information (how people share information in the group). Network sociologist Ronald Burt suggests [18] the enhancing of diversity by finding the gaps in the social network: individuals located near a network hole have better chances of having good ideas because opinion and behaviour are more homogeneous within the same group, therefore people connected across different groups are more familiar with alternative ways of thinking and behaving [19]. If people in the group share the same thinking and cultural patterns it is likely that they will find similar solutions, thus adding people with different mind-sets might help a group get unstuck by introducing new perspectives and heuristics.

The process of sharing information influences the diversity of a group since communicating can potentially exert group pressure for individuals to understand problems the same way as others. From a psychological point of view, individuals feel more comfortable conforming with the group because it drains less energy than the confrontation with other perspectives. This behaviour falls into the logic of Groupthink, a psycho-social phenomenon associated with peer pressure in groups to move towards conformity. The result of thinking conformity is that groups tend to use the same thinking strategies adopting an unproductive perspective, leading to similar solutions, which ultimately evolve into overall weaker solutions [13].

In a wise crowd, aggregated results of individuals solutions are largely dispersed but unbiased which allow for errors to cancel each other out. The notion that collective error is equal to the average individual error minus the diversity of a group is quantified in the Diversity Prediction Theorem (see Table 3) proposed by S. Page [13].

The first common sense logic that might be assumed when referring to collective wisdom is that *the more accurate individuals a group has, the more accurate the group will be*. But as we will see in page 11, that is imprecise because *diversity trumps ability*. The accuracy of a group is thus the result of individual accuracy and the group diversity.

$$\text{Collective Error} = \text{Individual accuracy} - \text{Diversity}$$



To understand the Diversity Prediction Theorem, the following simple example shows a prediction made by Alice, Jose and Pedro, on how many guests there will be this night at the El Farol Bar.

TABLE 1: AVERAGE OF PREDICTIONS

	<i>Prediction of the number of people</i>
Alice	42
Jose	56
Pedro	25
Average value	41
True value	39

The **individual error** captures how far individuals' estimates are from the true value. Errors are squared so that negative and positive errors do not cancel each other out [13]. In this case, as the true value was 39, Alice's prediction error can be calculated as  $(42-39)^2=9$ . Jose's error is  $(56-39)^2=289$  and Pedro's error is  $(25-39)^2=196$ . Alice has the most accurate estimate while Jose's estimate is the furthest from the true value of 39, so the most inaccurate estimate. The **average error** was  $164,6 = 9+289+196/3$ .

Now, to calculate how accurate the crowd was, **the collective error**, we calculate the square of the average prediction of the three people, which is 41, and the actual true value  $(41-39)^2 = 4$ .

Comparing both results, one can notice that the crowd's distance to the true value is less than any of the individual estimates, which means that the crowd was better at predicting than everybody in it. The crowd averaged estimate value was 41 while the best individual estimate was 9. This is the wisdom of the crowd effect.

The explanation for this asymmetry is found in diversity. Some individuals will make predictions too high and others will make predictions too low, which will allow for mistakes that partially cancel each other out, to become less severe [13]. The diversity of predictions is measured by averaging the values of each individual estimate's squared distance from the collective prediction. Following the previous example:

TABLE 2: EXPRESSION OF DIVERSITY

	<i>Prediction</i>	<i>Diversity</i>
Alice	42	$(42-41)^2=1$
Jose	56	$(56-41)^2=225$
Maria	25	$(25-41)^2=256$
Diversity	$= (1+225+256)/3=160,6$	

**Diversity** is the average square distance between everyone's estimates and the collective average estimation (also referred to as variance), which translates to how different individual estimates are and how they relate to the collective guess.

The Diversity Prediction Theorem is a mathematical identity that grasps the relationship between individual's accuracy, the diversity of estimates and how it affects collective error. So in order to have a small CE (or the WoC to occur), the AE needs to be very large, otherwise everyone could solve the problem fairly easily and consequently the crowd would not necessarily predict better than the individuals in it. Thus, if having a large AE is essential for the WoC to manifest, D will necessarily need to be large too, as we can see in Table 3:

TABLE 3: DIVERSITY PREDICTION THEOREM [13]

Collective Error (CE)	=	Average of Individual Errors (AE)	-	Diversity (D)
Collective Error is the distance between crowd consensus and the external truth.		Average distance between individual and external truth.		Distance between each individual response versus the crowd response. If the distance is zero represents no diversity.

EQUATION 1 - DIVERSITY PREDICTION THEOREM EQUATION [13]

$$(c - \theta)^2 = \frac{1}{n} \sum_{i=1}^n (s_i - \theta)^2 - \frac{1}{n} \sum_{i=1}^n (s_i - c)^2$$

Using the previous example to validate this theorem:

Collective error (CE)	=4
Average of individual error (AE)	=164,6
Diversity (D)	=160,6

So that:

$$4 = 164,6 - 160,6$$

For the wisdom of the crowds to exist, the CE value needs to be small, otherwise the crowd is not smart. Departing from the theorem, CE=AE-D, a *mad* crowd is a

crowd where the CE is very large. Consequently, the AE is also very large so D will inevitably need to be small. So the *madness of crowds* (high CE) comes from like-minded people (low diversity) who are most of the time wrong (high average error). In conclusion, the Diversity Prediction Theorem explains how essential diversity is to attain the WoC effect.

### 2.2.2. Deconstructing Cognitive Diversity

To explore the idea of diversity, it is necessary to look at the principles behind problem solving. Page [13] defines the four formal frameworks involved in the problem solving process:

(1) **Perspectives** are ways of representing situations and problems. Perspectives define how people interpret reality and organize their thinking with an internal language. It is a map from reality translated into language such that each object has a symbol. An illustration of what perspectives consist of are the different perspectives people have when organizing lists, for instance a real estate listing of houses: organized by area, location, number of rooms, or price. Different organizations will generate different landscapes with peaks, for example the house with the biggest square foot area. Good perspectives create landscapes with a single peak, meaning that the perspective organizes information in a meaningful way so that finding a solution for that problem becomes clearer.

Each problem solver can be characterized by their local optima (local peaks) and the probability attached. Better solutions depend on how diverse a landscape is. If someone brings a diverse perspective, then for each problem, each perspective creates a landscape where peaks are different from the average. New perspectives often are a result of previous perspectives combined.

(2) **Heuristics** are learned rules applied within perspectives to find solutions or a way of constructing solutions. Heuristics perform differently according to the type of problem, meaning that for each problem some heuristic will fit, and others will have generate bad results. To *deal with the bigger parts of a problem first* is an example of a heuristic. Some problem types might require this approach but other problems will not likely benefit from it.

(3) **Interpretations** are categorizations of reality. It is a mapping of objects, situations, problems and events into words, where one word can represent many objects.

(4) **Predictive models** are interpretations and predictions for each category of the interpretation, for example, "It seems like it's going to rain". In order to make a prediction it is necessary to have a way of representing the entities of the

prediction. Interpretations - categorizations - based on perspectives, combined with experience and theory, are the constructors of the predictive model.

### 2.2.3. Diversity trumps ability theorem

The Diversity Trumps Ability theorem assumes that a collection of average diverse problem solvers beats a collection of expert problem solvers. This is the very principle behind the wisdom of the crowds: collections of average individuals can have better results than a collection of experts. To demonstrate this logic, from a universe of  $N$  people we compare the collective performance of the  $M$  best problem solvers against the performance of a random collection of  $M$  problem solvers [13]. Diversity Trumps Ability theorem states that random collections outperform best collections.

The first condition for diversity to trump ability is that **the problem must be difficult**, meaning that no individual problem solver always locates the global optimum. Because difficulty lies in the eyes of the beholder, the difficulty of a problem is in relation to any of the problem solver's perspective. If the problem is so easy that any problem solver can always find the best solution, then the collection of the best individuals will always locate the solution, while a random group might not contain everyone who always finds an optimal solution. The second condition to support this theorem is that problem solvers must have access to some amount of information and all of them have to have some ability to solve the problem - the **Calculus Condition**. If individuals make their estimates at random because they lack information or the ability of reasoning, then collective estimation will carry no information. The WoC phenomena does not imply that individuals have the exact right values but that the ability of wrong guesses to cancel each other [20]. This also means that not every problem is suitable, for example asking a random collective to solve a mathematical problem might not produce the best results, while asking an expert - a mathematician - would most probably result in a better answer. Whereas if a mathematical problem is given to a diverse group of mathematicians, the collective solution will likely outperform the ones from a smaller expert group of mathematicians as it holds more diversity.

The third condition - **the Diversity Condition** - implies that any solution other than the global optimum is not a local optimum for some nonzero percentage of the users [13]. It means that given any non-optimal route, an individual has to be able to find an improvement without finding a solution.

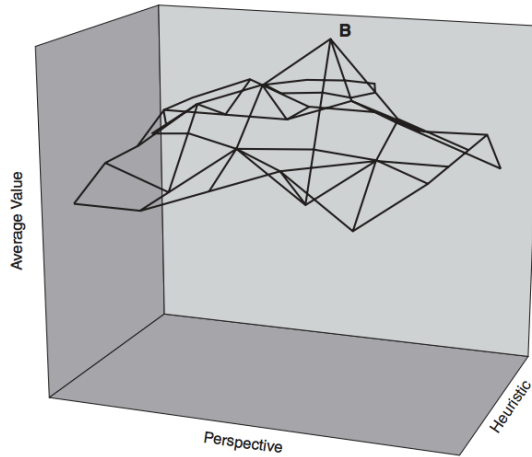
The fourth condition refers to the size of collection. **The population of problem solvers must be sufficiently large**. The more difficult problems are, the more local optima they have, thus more problem solvers are necessary to overcome the problem of overlapping local optima. There is no explicit size for a collection to solve a problem successfully; the exact number depends on the difficulty of the problem and the diversity of the collection.

Having all the four conditions - the problem must be hard, the collection of the problem solvers must be smart, the collection must be diverse and the group size large - is proven sufficiently to uphold the diversity trumps ability theorem [13].

Given the above mentioned conditions, a random selection of problem solvers outperforms a collection of the best individual problem solvers. We can represent in a diversity landscape by setting the height of a perspective/heuristic pair equal to its average value applied to a difficult problem. Each peak in the landscape represents a problem solver. The landscape has a global optimum representing the best individual problem solver. For difficult problems no single pair of perspective/heuristic locates the global optimum all the time. If we generate several problem solvers, the top problem solvers will gravitate around the global optimum  $\alpha$ , but on really hard problems even the globally optimal problem solvers can't find the global optimum to the problem [13].

FIGURE 1: HEURISTIC BOX

*Page's Perspective/Heuristic box [13]*



Considering Figure 1, the peak  $B$  has the best set of perspectives/heuristics for that specific problem. All the peaks situated around  $B$  represent problem solvers with similar perspectives and heuristics that will not perform much better as a collection than they do individually. By decreasing the number of problem solvers, the landscape gets scattered pairs of perspectives/heuristics and the clustering around  $\alpha$  becomes imperceptible.

Diversity of perspectives/heuristics allow problem solvers to perform better collectively than individually, but alone, are sufficient to outperform the best. To do so, the collection must be smart (second condition), the collection must be large (fourth condition), problem solvers cannot get stuck at the same group of solutions with low values (third condition) and the problem must be difficult (first condition) [13].

## 2.3. Measures

### 2.3.1. Aggregation methods

The aggregation method is used to extract value from the crowd. The method used can vary from simple arithmetic averages to more complex methods that combine several aspects of the estimates. As reported in many experiments, the chosen method deeply influences the final *wisdom* value. To create an aggregated estimate, collective data must be first gathered, centralized, and processed with a quantitative method, to achieve the final value, which is then measured in terms of accuracy by comparing it to an external truth. The principle behind the WoC effect is that individual knowledge can be extracted by eliminating subjacent misinformation, thus the success of the crowd's prediction will highly depend on how an aggregation model is applied, specific conditions of the crowd (as seen in the previous section) and the nature of the problem.

#### ARITHMETIC MEAN, MODE, MEDIAN

In his weight-judging competition experiment [1], Galton used an arithmetic mean which successfully predicted the correct answer within 1% accuracy. Currently, most experiments use the arithmetic mean, median and mode as starting points to analyse data and then, according to the type of data, fine tune their aggregation methods by complementing it with weighed mean or more complex models.

#### GEOMETRIC MEAN

Lorenz et al. have conducted an experiment [2] in which the crowd effect exists with respect to the geometric mean but not the arithmetic mean. The study reports that only in 21,3% of the cases is the arithmetic mean closer to the truth than the individual first estimates, stating that the type of questions is at the origin of non-normally distributed estimates where the majority of estimates is low and a minority is scattered in a fat right tail (as log-normal distributions are). Participants were asked to estimate real world geographical facts and crime statistics for which they were unlikely to know the right answer but at the same time had some knowledge about, for example: "What is the population density in Switzerland in inhabitants per square kilometre?". Because participants had difficulty in choosing the right magnitude of their estimates they faced a problem of logarithmic nature. In this study, the geometric mean (exponential of the mean of the logarithmized data) performs better as an aggregator method because when using logarithms of estimates the arithmetic mean is closer to the logarithm of the truth than the individuals' estimates in 77,1% of the times [2].

## DECISION MAKING MODELS

In a study that investigates the wisdom of the crowds in the TV show *The Price is Right* [21], decision models have been used to understand if the knowledge of all four participants can be combined to provide a good estimate of the value of the prize.

Because the estimation of the price happens in a competitive context, participants don't necessarily bid what they believe the correct price is, instead they play strategically. For example, four players place bids of \$650\$, 675\$, 110\$ and 1\$ to win a 960\$ stereo. The second bidder would be the winner because it was the closest to the true price without exceeding it. Therefore, to assess the WoC effect, averaging the bids is not necessarily the best way to combine the knowledge. Instead, combining the knowledge about the prices that led to the bids provides a much more useful insight: averaging what they know, not what they say.

The sequences of bids from 72 competitions were analysed and measured against 11 aggregation methods - from which four were based on decision models - using the mean absolute error and the mean relative error between to measure the distance between the aggregate estimate and the true price. The first quantified how many dollars away from the true price each estimate was on average. The second quantified the proportion of the true price that an estimate differs from the true price on average [21].

The 4 decision models try to capture the strategy behind the bids: the first model assumes that all players choose between bounds non-strategically, so for the example above, the estimate would be 551\$, the median value between 1\$ and 1100\$; the second model is a natural adaptation of the Thurstonian model and also assumes that players played non-strategically from their range but allowed individual differences; in the third model, the first three players choose between bounds non-strategically but the last player bids according to the probability of winning and uses the inferred mean between the bounds. The fourth model also assumes that the two players choose between bounds non-strategically but the last two bid according to their probability of winning, and uses the inferred mean between the bounds as the estimate [21]. The decision model where the two last players bid strategically provided the best performance, followed by the model where only the bid of the last player was strategic. Moreover, the best six methods all involved aggregation denoting the wisdom of the crowd effect.

## LOCAL DECOMPOSITION METHOD

Solving multidimensional problem-solving tasks require the combination of local aspects of solutions into a global solution. For problems such as the Minimum Spanning Tree (MST) *local decomposition method* allows the problem to be broken down into common pieces. In this case, it is expected that good local connections between nodes are more frequent than bad solutions. As a result, the best answer to the problem will be a collective agreement between specific connections between

nodes using an agreement matrix where values are transformed into a cost matrix such that edges with higher agreement are given lower costs [14].

#### GLOBAL SIMILARITY AGGREGATION METHOD

Other approaches try to capture the whole solution not by decomposing the solution into parts, but by finding the individual solution that is most similar to other individual's responses. This method does select the prototypical solution and it cannot therefore identify new solutions, so resulting aggregations can never generate better responses as the individual solutions. Similarity is calculated by the proportion of coincident edges among individuals [14].

#### 2.3.2. Considerations about aggregation

The above methods of aggregation consist essentially of averaged estimates of individuals externally calculated. Contrary to other nature occurring phenomena that involve averaging, such as path formation (or any other stigmergy phenomena), the wisdom of crowds as observed in the literature is not a direct consequence of local interactions, but instead a centralized computation of individuals' estimates.

#### 2.3.3. Assessing the strength of WoC effect: Wisdom of the crowd indicator

The wisdom of the crowd indicator [2] is based on quantile statistics to measure how many central estimates (in the ordered sample) are needed to bracket the true value, so a group is considered to be maximally wise if the truth lies between the two most central values of all estimates, to which is attributed the highest WoC indicator value. The indicator is denoted as follows  $\{i | \hat{x}_i \leq \text{truth} \leq \hat{x}_{n-i+1}\}$  and it achieves its maximum at  $n/2$  when the truth lies between the two central values implying a higher WoC when truth is closer to the median. The WoC indicator is an attempt to quantify the quality of the WoC effect, assuming that in two groups with the same collective error, the group with large dispersion shows more WoC than the one with small dispersion because the crowd does not outperform the individuals by a large magnitude.



## 2.4. Typology of problems

Wisdom of the crowds has been successfully applied to a large set of different problems from finance, politics, computer science to military operations. From the reviewed literature it is possible to outline three main categories frequently mentioned for the types of problems solved by the WoC: 1) estimation of present or future events 2) problem solving and 3) decision-making. However, there is no precise source in which to base these conceptual distinctions, thus in order to understand conceptually the range of tasks that can be solved by groups and how the WoC can be applied to each task type, we use as reference the task classification scheme from McGrath, a well-established taxonomy of group tasks based on the type of coordination process they require [22].

FIGURE 2: MCGRATH'S CIRCUMPLEX MODEL

*McGrath Circumplex model for group task types [23]*

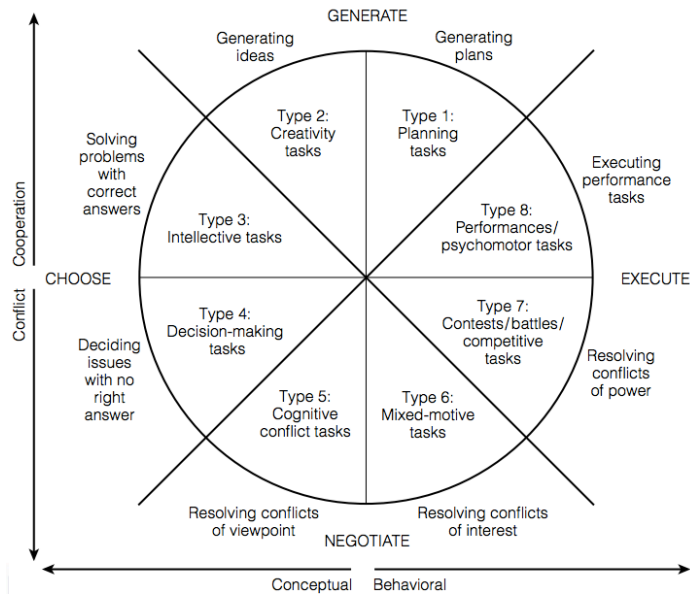


TABLE 4: MCGRATHS QUADRANTS

QUADRANT I	GENERATE	
Type 1	Planning tasks: Generation plans, <b>Key notion:</b> Action-oriented plan	Cooperation
Type 2	Creativity tasks: Generation of ideas <b>Key notion:</b> Creativity	

QUADRANT II		CHOOSE
Type 3	Intellective tasks: solving problems with a correct answer. Logic problem and other problem-solving tasks with correct but not compelling answers. <b>Key notion:</b> Correct answer	Abstract reasoning tests
Type 4	Decision making tasks: Dealing with tasks for which the preferred or agreed upon answer is the correct, choice, juries, no right answer. <b>Key notion:</b> Preferred question	Voting
QUADRANT III		NEGOTIATE
Type 5	Cognitive conflict tasks: resolving conflicts of viewpoint. <b>Key notion:</b> Resolving policy conflicts	
Type 6	Mixed motive tasks: Resolving conflicts of motive of interest. E.g., negotiations and bargaining tasks mixed motive dilemma tasks. Coalition formation rewarding allocation tasks. <b>Key notion:</b> Resolving pay-off conflicts	
QUADRANT IV		EXECUTE
Type 7	<i>Contests/Battles</i> : Resolving conflicts of power; competing for victory. E.g.: wars, all winner-take-all conflicts, and competitive sports. <b>Key notion:</b> Winning.	
Type 8	<i>Performances</i> : Psychomotor tasks performed against objective or absolute standards of excellence, e.g., many physical tasks; some sports events. <b>Key notion:</b> Excelling.	Typing test: a group is asked to type a Wikipedia article within limited time. (2 <sup>nd</sup> best predictor of g factor)[22]

Further in this chapter we present a list of WoC application examples (**Error! Reference source not found.** on page **Error! Bookmark not defined.**), listing all known experiments from the present literature on the subject. By comparing the McGrath task classification taxonomy with the gathered examples, we derive that all existing examples of WoC fall onto *Quadrant II – Choose*, and from these nearly all of them fall onto Type 3 - *Intellective tasks with a correct answer*.

Following this model, Intellective tasks (Type 3) comprehend any problem solving or estimation of events (present or future) where a truth-value is given at any point, whereas Decision Making tasks (Type 4) correspond to decision making tasks where the expressed preference or agreement is the outcome with no truth-value, typically involved in voting processes. Although this model does not specify the differences between problem solving tasks and estimation tasks (both regarded as intellective tasks with a right answer), it contextualizes these tasks by providing an overall notion of the tasks solved by groups. Secondly, it casts light on the potential to explore other quadrants, which remain quite unknown to the effect of WoC.

Next, we draw examples from the reviewed literature to illustrate the problem typology segmentation we propose: estimation of present or future events and problem solving for single and multiple variable (decision-making).

#### **2.4.1. Estimation of events**

Estimation of events constitutes by far the most representative application of WoC. Events can vary in time: asking a number of people to estimate the number of jelly beans in a jar or who was the first president of the European commission are examples of the estimation of present events, since the truth-value is presently available; if otherwise, the truth-value is only available at a given moment in the future, the task is therefore an estimation of a future event or more commonly referred to as a prediction.

##### **PRESENT EVENTS: MISSING SUBMARINE**

In May 1968, the U.S. submarine Scorpion disappeared. The Navy knew nothing about what had happened to the submarine, only the Scorpion's last reported location, and that the destination should have been Newport News Virginia. The Navy started a 20-mile wide search in water thousands of feet deep, a process that would eventually take many months to complete. However, a naval officer, John Craven, had a different plan as he concocted a series of scenarios about what might have happened to the submarine and assembled a team of men with a wide range of knowledge including mathematicians, submarine specialists and lay men. He didn't create a round table but instead asked each one individually to give their best estimation of the location of the submarine. Then he used Bayes's Theorem to average the estimates out. The formula calculated how new information about an event changes the pre-existing expectations of how likely the event was [24]. Five months later the submarine was found 200 meters away from the estimation of the group.

##### **FUTURE EVENTS: PREDICTION MARKETS**

Prediction markets are speculative markets, similar to stock markets, where predictions can be extrapolated from the exchange of stocks and market fluctuation. The first prediction market was founded in 1988 during the U.S. presidential campaign. The Iowa Electronic Markets (IEM) run the by the College of Business at the University of Iowa, is open to public participation and allows participants to buy and sell futures "contracts" based on their predictions. The IEM is a real money market with a maximal set on 500 dollars and the average user has 50 dollars at stake. There are several kinds of contracts, but two contract types are the most common. The first one is designed to predict the winner of an election. If a participant thinks candidate A is likely to win she buys a "Presidential candidate A to win". The price each contract has reflects the

market's judgement for the win probability. If the prediction is correct a monetary reward is given, otherwise, none. The second type of contract is to predict what the percentage of the final vote each candidate will have. The payoffs are determined and proportional by the vote percentage.

A study to assess the performance of IEM in forty-nine elections between 1988 and 2000 has shown that the IEM has generally outperformed the major pools and is able to be more accurate even months in advance of an election. Between 1988 and 2000, 596 pools were released. The IEM's stock price for each day was 3/4 of the time more accurate.

Other similar markets are for example the Hollywood Stock Exchange (HSX) that uses estimates to predict the outcome of the Oscars winners. Its accuracy is not as high as the IEM likely because the wagering is entirely done with play money. Status, reputation and incentives encourage the investment of energy and time to provide the best estimates [24].

#### **2.4.2. Problem solving: single variable**

Problem solving tasks can be similar to estimation tasks, but require some degree of involvement in reaching a solution and a greater access to information in order to solve a problem.

#### **PROBLEM SOLVING: SOLVING COMPLEX PROBLEMS**

One study where the WoC has been applied to problem solving tasks was on the Traveling Salesman problem and Minimum Spanning Tree problem [14], which suggested that this effect can also be observed for problems that demand coordination of multiple pieces of information. The Traveling Salesman (TS) problem and Minimum Spanning Tree (MST) problem are two well-known computer science multidimensional problems for which there are near optimal solution paths algorithms. The aggregation models were developed in order to optimize the relation between individual and group responses that have demonstrated a strong wisdom of the crowd effect. The first aggregation method developed divides the MST problem into smaller pieces and then combines the common parts of individuals' solutions into a global solution considering collective agreements on particular edges are better aggregate solutions. Solutions selected by aggregation models perform better than sole individuals' solutions, either by performing better than the best individual's average, or by exceeding the vast majority of individuals.

#### **2.4.3. Decision making: multivariable**

Decision-making (multiple variable problem solving) is always part of the process of solving a problem (finding the best outcome) but involving some degree of cooperation between the parts by expressing a preference. Some problems highlight the feature of cooperation more than others. For example Wikipedia, is a

collaborative and cooperative platform that contains a decision-making back-office tool in which authors discuss possible editions to an article. It is through this discussion (which could coarsely be comparable to an aggregation method) that a *better* solution for that article can be found.

#### DECISION-MAKING: GOVERNANCE

In 2007 the New Zealand government, in an initiative to incentivize citizen participation, launched an open wiki for public edition of the 50-year old New Zealand Policing Act. Citizens were able to express their ideas in the wiki to shape the new Policing act. While this example cannot be considered to be under a WoC, it does set the initial mechanism to enable collective decision-making [9].

TABLE 5: COMPREHENSIVE WOC LITERATURE REVIEW

Problem solved collectively	Problem solving type	Nature	Output	Quantifiable?	Ground truth (correct answer)	Aggregation method	WoC with Social influence	Social influence inference	Collective outperforms experts?	Size of the crowd
Marbles in a jar, temperature in the future[20]	Type 3	Estimation	Numeric	Yes	Yes	Mean	No	-	Yes	30
Marbles in a jar [25]	Type 3	Estimation	Numeric	Yes	Yes	Median	<b>Yes</b>		Yes	82(429)
Prediction Markets	Type 3	Forecasting	Numeric	Yes	No	Majority	No		-	Yes
New Zealand Policing act [9]	Type 4	Decision making	Preference	No	No	*Local human decision	No		Yes	-
General knowledge testing [2]	Type 3	Estimation	Right answer	Yes	Yes	*Geometric mean, arithmetic mean Median	No		Yes	144 (12 groups)
Cultural Market [3]	Type 4	Decision making	Preference	Yes	Yes (arguably)	Majority	No	Number of downloads)	-	14341
Ox weight contest [1]	Type 3	Estimation	Numeric	Yes	Yes	Mean	No	-	Yes	800
Who wants to be a millionaire's audience	Type 3	Guessing	Right answer	Yes	Yes	Majority	No	-	Yes	~80
The WoC with communication [26]	Type 3	Estimation	Rank order	Yes	Yes	Average Kendal's Tau Distance Borda count method	<b>Yes</b>	By providing last person's estimate	-	172
The price is right [21]	Type 3	Decision making (competitive and strategic)	Right answer	Yes	Yes	*Average of the middlemost two bids; random bid, non-strategic average (other 11)	<b>Yes</b>		-	4 show participants
WoC in one mind [27]	Type 3	Estimation	Right answer	Yes	Yes	Geometric mean	N/A		-	144 (12 sessions)
The crowd within [28]	Type 3	Estimation	Right answer	Yes	Yes	Mean; Squared mean	N/A			428

Peer to Patent [9]	Type 3	Crowdsourcing	Input to legal decision making process	No (forum based)	No	Local human decision	N/A		Yes	N/A
Web of trust[9]	Type 3	Crowdsourcing	Reputation	Yes	No	Rating	N/A		Yes (?)	N/A
Ordering problems associated with memory (Rank ordering) [16]	Type 3	-	Ranking	Yes	Yes	Bayesian version of a Thurstonian model	<b>Yes</b>		Yes	172
Travelling Sales Man [14]	Type 3	Estimation	Path	Yes	Yes (optimal solution)	Local decomposition model	N/A			101
Minimum Span Three [14]	Type 3	Estimation	Path	Yes	Yes (optimal solution)	Local decomposition model	N/A			
Voting	Type 4	Voting	Numeric	Yes	No	Majority	Yes and No		N/A	N/A
Missing submarine finding [24]	Type 3	Guessing	Spatial	Yes	Yes	Spatial mean	N/A		Yes	-
Wikipedia	Type 4	Collaborative	Textual	No	No (or to a certain extend yes)	Human (through discussion)	N/A	N/A	Yes	N/A
Amazon's recommendations	Type 4	Crowdsourcing	Recommendation (based on a preference)	Yes	Yes (a preference)	Rating	No		N/A	N/A

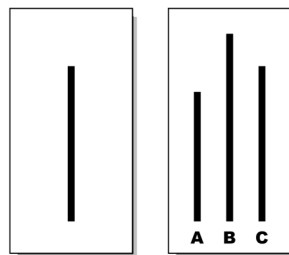
## 2.5. Social influence

### 2.5.1. Context

In the early 50's the psychologist Salomon Asch set out a number of experiments to understand how social forces affect individual opinions and attitudes. The conformity study meant to test whether individuals could change other's judgment of a situation without changing their knowledge or assumptions of that situation. Asch gathered 9 male students in one room, all confederates but one. The study began with the experimenter showing two cards, the first with one single line, and the second with three lines of differing lengths. When participants were asked to announce their answer to which of the lines in the second card was identical in length to the first one, the confederates gave unanimously incorrect answers 12 out of the 18 trials. The one participant did not know that all the others were confederates as they also gave correct answers so he would not suspect of collusion.

FIGURE 3: CONFORMITY EXPERIMENT

*Pair of cards shown in the Conformity experiment (card 1 and 2 respectively).*



After testing 123 young men Asch discovered that when alone, individuals would make mistakes less than 1% of the time, whereas in the social setting he had created, individuals tend to make errors in judgement 36,8% of the time. Three out of four people gave incorrect answers to very simple questions after hearing other answers in the group. In the interviews after the experiment, the conformists gave explanations like “I am wrong, they are right”, “not to spoil the results”, or even had the perception others were engaging in herd behaviour without noticing that they were conforming too, underestimating the frequency of their conformity.

What this study reveals is that in a group, opinions and attitudes can be highly sensitive to the group's majority view. Now if we think of groups where individuals have grown strong ties between them or have similar perspectives and backgrounds, the tendency will be to converge even more strongly towards group conformity. This notion is important to understand how social influence can modify the attitudes, opinions and judgements of individuals in groups even if solving very simple tasks.



### 2.5.2. Social influence impact on the wisdom of the crowd effect

Social influence is broadly suggested in literature as an inhibitor the WoC. Several studies [2][3][29] point out that under social circumstances the wise effect of the crowd dissipates as diversity narrows down. Apparently, having access to the estimates of others prompts individuals to revise and adjust their estimates to the ones of the group converging gradually into a consensual biased point - based on the assumption that others might have better estimates, more information, or merely because individuals feel prone to follow the crowd, have peer pressure toward conformity [2] or adopt a group strategy.

The social influence effect is described in [13][2] as a statistical effect that undermines the wisdom of the crowds by decreasing the diversity of groups without improving their accuracy. Estimates of individuals will tend to converge at some point due to influence of the group and become biased towards a wrong value. Another statistical consequence of social interaction is the Range Reduction Effect [2]. If all estimates of a group are narrowly distributed around an incorrect value, any subsequent estimate would gain confidence to produce a wrong estimate. To illustrate this concept, the landscape on page 12, shows the distribution of estimation values for a given problem. If the value-answers for that problem gravitate around an area outside of the truth region, then, any posterior estimate would be biased to be located anywhere near that area. A good indicator of the WoC, which generalises the concept of bracketing the truth, considers a group to be maximally wise if the truth lies between the two most central values of all estimates. On other side, the psychological consequence of the two mentioned effects, is the Confidence Effect. Individual's confidence boosts when social interactions allow for their estimates to gain more acceptance [2].

A study about a cultural market of music [3] demonstrates that by showing the number of downloads next to a song - a form of social influence - any average quality song can become a hit because quality is perceived as popularity. They have parameterized songs according to its quality, and found out that in the group where there was no indication of the number of downloads, the most voted songs matched the best songs, in the group where social influence was present in the form of the number of downloads, hence the popularity of a song, average quality songs became the most voted. Quality also determined partly the success of a song: while the best songs rarely were of poor quality, the worst songs rarely did well, but anything in the interval was possible. The introduction of social information determined the unpredictability and inequality of success. Groups can be remarkably accurate in estimating facts or solving problems with little knowledge about the problem. As noted previously, one fundamental condition of the crowd to be wise is diversity, which implies that individuals are independent from each other. Because social influence affects how individuals take decisions and make estimations, it impacts directly on the degree of independency of estimates of the crowd and consequently decreases diversity.

Particularly one study [2], has been used as reference for methods and

comparative analysis for our data. In this study, 144 participants were recruited to participate in estimation tasks testing their real-world knowledge such as “How many murders were officially registered in Switzerland in 2006?”. Twelve sessions took place, each one testing twelve participants at a time. For every question, participants were elicited to give a consecutive answer 5 times and rate the first and fifth response with a six-point Likert scale (1, very uncertain, 6, very certain).

To test the impact of social influence, three different information conditions were tested for each round, subjects could base their second, third, fourth, and fifth estimate on: “aggregated information”, in which subjects received the average (arithmetic mean) of all 12 estimates of the former round; “full information”, in which subjects received a figure of trajectories of all subjects over all previous rounds; the “no information” treatment served as control group where no group information was shown.

The results showed that the arithmetic mean performed poorly. Adjusting the aggregation formula to the geometric mean (the exponential of the mean of the logarithmized data) provided slightly better results - a 11,9% distance to the true value is the most successful result. According to the authors, the evidence for a social influence effect lies in the statistical tests (Kolmogorov–Smirnov, f-tests and t-tests) performed showing that the group diversity is significantly reduced under social influence, whereas the collective error changes only slightly [2]. Another aspect of the WoC introduced in this study is the concept of *range reduction effect*, which translates into the idea of “bracketing the truth”: estimates narrowly distributed around the wrong value will deliver the wrong hint regarding the location of the truth, and perhaps even gain more confidence if a dense clustering forms around the wrong value. To quantify this, the wisdom of the crowd indicator considers a group to be maximally wise if the truth lies between the two most central values of all estimates: a high wisdom of the crowd implies that the truth is close to the median, implicitly defining the median as the appropriate measure of aggregation.

### 2.5.3. Social interaction as an aggregation mechanism

As we have seen, individual decision-making processes are susceptible to social influence. However, only artificial conditions allow for social influence to be selectively isolated. Because people exist within social networks and communication is a fundamental part of being human, it is an impossible task to put a barrier up to where social influence starts and ends. Even the smallest amounts of social information can change how individuals take decisions and develop into herding behaviour [2].

At the same time, other collective intelligent phenomena seem to use social interaction as means of spreading innovation, or more objectively, the best available answer to a specific problem. For example, considering collective behaviour, navigation accuracy in humans and animals benefits from a large

number of individuals, where the average over each other's directional preferences takes the group towards the right direction by cancelling individual directional errors that decreases as a non-linear function of group size [11].

Quorum decision-making in social animals also suggests that a few knowledgeable individuals with different information can compete between each other by fomenting local interactions and can generate collective decision-making towards the best available decision. *Temnothorax Albipennis*, a species of small ant colonies, move their nest sites frequently. Because the size of colonies is rather small, pheromone communication is not effective as the capacity for pheromone trail reinforcement is insufficient. How can they achieve consensus when faced with multiple possibilities? When a colony needs a new nest, approximately 30% of ants scout for new sites using visual cues. Each ant assesses the new site, taking into consideration, among other properties, size, entrance size, and brightness. If a site is positively perceived, the search is discontinued and the ant returns to the current nest where she will start recruiting other single individuals. New ants will follow her closely behind and learn the route. Upon reaching the nest, she will evaluate the nest independently and if the quality of the site is also positive, she will also become a recruiter. If recruiting ants detect a threshold quorum of ants present in the new nest, they physically start to carry each other from the old nest to the new, rather than to lead them. Amplification of recruitment to one site inhibits transport to other sites because there are less potential scouts willing to move to other nests [30]. Similarly, honeybees also recruit other bees to assess new sites, but instead they perform a waggle dance to inform others of the direction of their find, the dance length being proportional to the quality of the perceived site. Probabilistically, positive feedback will make more ants and bees recruited to better sites.

Quorum decisions have the advantage of enabling multiple comparisons between options based on individual information where the risk of copying cascades of inexact decisions is unlikely because it would take several individuals to come to the same conclusion independently to reach the same quorum threshold [11].

In contradiction with Rational Choice Theory, that states that each individual makes rational decisions in order to maximize fitness, violations of rationality have been repeatedly observed in animals and humans [31]. The assumption that natural selection shapes decision-making in order to attain the highest individual profit finds obstacles at the highest level of observation. To assess rationality is necessary to look at the adherence of consistency principles: a preference for choice A over B should not change by introducing a choice C. *Independence from irrelevant alternatives* refers to the insusceptibility of a decision in relation to decoys. Animal decision-making occurs under strong constraints: time, cognitive limits and incomplete information that selects for heuristics (economizing computation by either excluding information or processing imperfectly). Sakaki and Pratt [31] tested the decision-making and recruitment processes of ants (*Temnothorax*) during nest-site selection and concluded that collective decision-making can eliminate

irrational errors of sole individuals by suppressing systematic errors that emerge from decision heuristics from cognitively limited individuals. In this sense, collective decisions compensate individual error, which is more likely to occur because single individuals have less information than groups of individuals

#### **2.5.4. The impact of imitation**

Crowd panic, riots, fads, mobs, fashions, all these are examples of social influence in a crowd at the level of behaviour, where small signals are amplified by imitation and become a large scale phenomenon. The most common manifestation of social influence is imitation, which is a positive feedback mechanism that plays an important role in the spreading and dissemination of innovation across a community, while economizing cognitive resources.

Imitation is a sophisticated skill that requires advanced cognitive skills - true imitation has been reported to appear almost exclusively in humans [7]. When allied to diversity and adaption, imitation is one of the most successful methods to learn effectively. The pressure to conform can have several reasons, but the most common comes from the desire of people to obtain social approval, who tend to have less probability to conform with the group if estimates are private, minimizing influence. However, the extension of conformity is sometimes deeper, and people still conform with the group even if responding privately [32].

So far, the majority of experiments confirm the wise effect only in crowds where the independence of guesses is artificially achieved, not allowing individuals to be socially contaminated by each other, therefore keeping the diversity of the group. One of the strongest criticisms aimed at this sort of experiment, is the restricted resemblance to real world situations, where individuals possessing useful information will most likely use it. A person at a bookshop might use the fact that a book has reached one million copies sales to decide whether to buy it or not. It doesn't mean that he or she is always using the same heuristic of buying what others bought, but the fact that the information is available might be useful for example depending on other variables such as time restriction. Humans are social animals and most activities have some degree of sociability (education, work or sports) thus it is nearly impossible to quantify social influence and define clear boundaries of independency.

We can conclude from the literature review introduced here, that collective intelligence phenomena, more concretely the wisdom of the crowds, is still an infant field with some incoherencies and aspects to be explored. Particularly regarding social influence, there is a consensual view that it interferes with the wise effect but it is limited to very few experiments at the time when we carried out the literature review. Some of these experiments have had positive results when inhibiting social influence but others have reported less obvious outcomes.

Based on the notions introduced in this chapter around the psychological aspects of social influence and quorum decision processes entailing imitation observed in human and animal collective behaviour, we have designed an experiment to assess the degree to which social influence impacts the WoC and more specifically, the degree to which the access to more information benefits the WoC by establishing multiple comparisons between the groups estimates and individual information, whose methods and design we will explain in the next chapter.

## 3. Methods

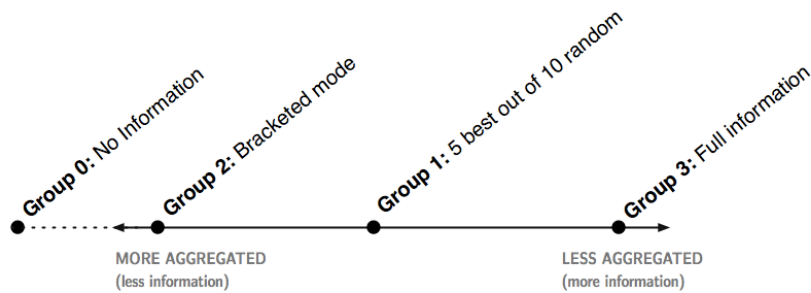
### 3.1. Experiment Design

#### 3.1.1. Introduction

To test the impact of information on the WoC effect we set up four groups where participants had access to different degrees of information. Besides the control group, which entailed no information sharing, the remaining groups all presented different degrees of information aggregation, i.e., the degree to which the information displayed had been processed onto a convergent value showed as a hint. The method to aggregate the estimates varies on how well the data represents the actual estimates. For example, participants on group 3 had access to a non-aggregated full information hint where every previous guess was represented by a value-dot – hence without any previous treatment - whereas on group 2 the hint showed an interval of two values where previous guesses had been more frequent, similar to a mode. In the latter case, by showing the compressed data interval, not only did we provide less information but also assumedly increased the likelihood of clustering around specific range of values, whereas by disclosing full information we intended to allow subjects to make an internal inference based on the estimates of others.

FIGURE 4: INFORMATION DEGREES IN GROUPS

From left to right, figure 4 shows the groups in relation to information aggregation.



Moreover, groups also differ on how hints relate to the true value. Groups 2 and 3 display hints that have no relation to the truth, consisting in the absolute values of estimates, whereas the hint provided in group 1 entails a degree of quality of estimates by presenting the *best* (which is relative to the true value) estimates out of a random set of estimates.

To measure the results of our experiment, we first start by assessing the normality of our data using a Shapiro-Wilk test to define the appropriate measure for the mean. Then we measured collective accuracy by analysing the central

tendency of data and calculated the geometric mean, median, standard deviation and plotted the data distribution in percentiles for overview. Secondly, we measured the dispersion and accuracy of our data with the Wisdom of Crowd Indicator as specified in [2] to compare the distribution of guesses around the true value. We particularly compared our central tendency and WoC Indicator results with [2].

Additionally, to compare group performance of our groups, we used the measures involved in the Diversity Prediction Theorem (Collective Error, Average Individual Error and Diversity). Further, we tested if groups' diversity of estimates differ significantly (Levene's test) and if the difference between the medians is significant (Mann-Whitney test).

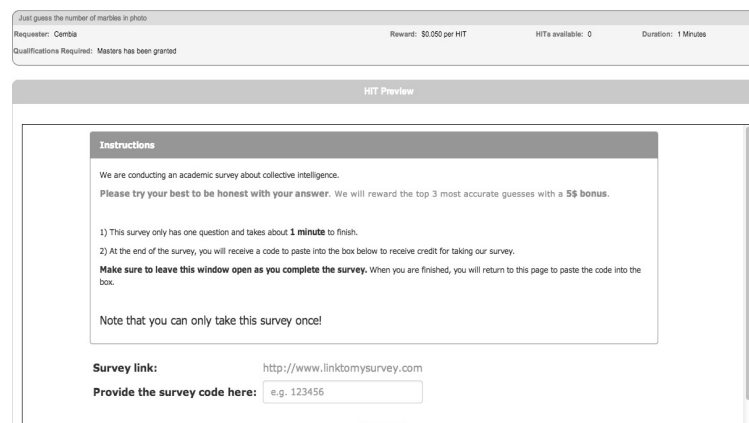
### 3.1.2. Set up of the experiment

The study used Amazon MTurk platform to recruit participants, manage payments, and direct participants to the survey page hosted on the Sciences Faculty of the University of Lisbon's server. MTurk workers, who choose to be assigned to this task, land on the experiment page on MTurk platform, which contains the instructions and a link to the survey. We have set the total amount of participants to 380 with a compensation of 7-dollar cents per assignment completion. The timer of the task was set to a maximum of 6 minutes (to avoid users from extreme rationality) and the entire batch was complete within 15 days.

A monetary incentive was promoted: 10\$ for the 3 estimates closer to the true value of jellybeans in the jar. We used a glass jar filled with 3150 jelly beans which we counted three times. The photographs tried to convey a perception as close as possible from the human eye and we used a frontal as well as a top view to offer sufficient spatial information, additionally using a clothes peg as a reference for scale.

FIGURE 5: EXPERIMENT LANDING PAGE

*View of the landing page in MTurk with a link for the survey.*



Participants were directed to the survey page once they clicked on the link hosted on MTurk platform and each one was randomly assigned to one of the four groups which varied only on the type of complementary information displayed as *hints*: a) no information, where no hint was shown b) shows 5 best out of 10 random of all estimates so far c) shows bracketed mode of all estimates so far d) dispersion map of all estimates so far.

Participants were asked to estimate how many jellybeans exist in the jar. The answer is a required field and must be a whole number. When participants submitted their estimates, the value of the estimation was stored in a text file along with the user's ID. For the information groups (1,2,3), the previous stored values were processed and displayed as part of the *hint* for the next participant.

FIGURE 6: SURVEY SCREEN

*Partial view of the survey screen (see: Appendix F to I)*

#### GROUP 0: CONTROL GROUP - NO INFORMATION

Participants falling into group 0 saw no additional information, therefore their estimates can be considered to be socially independent and unbiased. This information condition has been shown successfully in several studies [25][3][20] and it was used as the control group.

#### GROUP 1: THE 5 BEST OUT OF 10 RANDOM ESTIMATES

In this group we showed participants the five closest values to the true value out of ten random estimates. For every new estimate, a participant had access to an updated information hint, for example:

*Based on all the guesses of other participants, the closest guesses so far are  
(in no particular order) 3212, 3245, 3531, 3221, 3522.*



With this group we intended to test whether random best guesses would perform well as part of social influence, specifically the interplay between random values and the qualitative reference of *best* estimates, which provides relative information to the truth-value.

As the hints displayed were dynamical, refreshing with every new estimate, the first participant had no information displayed. To the first participant no information was shown. To the 2nd participant it showed one guess (the previous), to the 3rd it showed the two previous guesses and so on until the 5th participant. When the number of guesses was more than 5 but less than 10, participants 6th to 11th were shown a random selection of 5 guesses from the previous guesses. When the number of guesses reached 10, it chose the 10 closest guesses and randomly showed 5 of those to the participant.

#### GROUP 2: BRACKETED MODE

Participants in this group had access a bracketed mode of other participants estimates. We departed from the notion that showing the most frequent estimates could have a positive influence on crowd performance, but due to the large range of values we proceeded to a bin segmentation resulting in the following hint:

*Based on all the guesses of other participants, the guesses between 1000 and 2500 were the most common.*

The interval of the bin displayed is calculated by dividing, at each iteration, the range between the minimum and maximum value of all estimates in 10. So for example, if the estimates so far were 1000, 2000, 2000, 2500, 2500, the minimum and maximum values were respectively 1000 and 2500. This range is split up in 10 bins, which results in a 150 bin size starting from 1000 till 2500:

1000 1150 1300 1450 1600 1750 1900 2050 2200 2350 2500

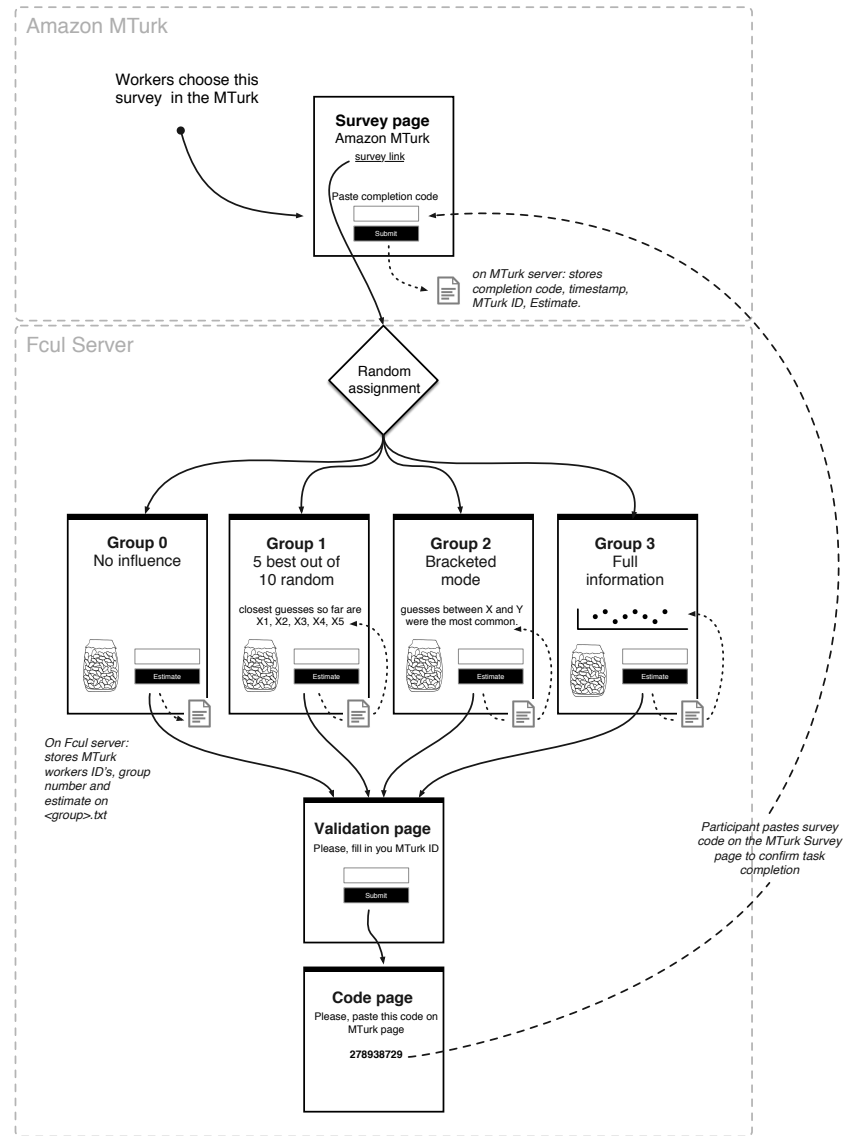
In this example, we have 1 value falling onto the 1000-1150, 2 values in the 1900-2050 bin, and 2 values on the 2350-2500 bin. The algorithm counts the number of estimates that fall onto a bin and picks up the one with the higher count. In case two bins have the exact same count of estimates the algorithm randomly selects one bin. The first two participants will not have access to the hint as it needs to build up upon the first and second estimates, therefore only after the 3rd participant hint the is shown.

### GROUP 3: FULL INFORMATION

In this group, participants had access to full information: a dispersion map with the values of all estimates so far with each estimate represented by a blue dot, which depicted identical or close estimations. The intention was to show the full range of estimations without any aggregated information, as this would assumedly be the most *natural* state of social influence. The full information state allows participants a vaster perception of the distribution of estimates, which calls for individual-based judgement based on multiple comparisons with others estimates.

FIGURE 7: EXPERIMENT FLOW

Participants were assigned to groups randomly. After submitting their estimate, they were directed to a validation page, which required the MTurk ID number; only then a completion code was displayed to paste on to the survey page in Amazon MTurk platform. Payments were managed on the MTurk platform.



## 3.2. Collection and treatment of data

### 3.2.1. Data collection platform: Amazon Mechanical Turk

The Amazon Mechanical Turk (MTurk) is an online crowdsourcing marketplace that allows individuals or companies (*requesters*) to distribute tasks that can be performed by any registered worker (*workers, turks, turkers*). A requester places a task in the MTurk interface - often referred as HIT's (*Human Intelligence Tasks*) - and defines duration and a price for the completion of the task, usually ranging from as little as \$0.01 for simple tasks, such as tagging an image, up to a few dollars for more involved jobs, such transcribing audio clips or writing product recommendations for ecommerce websites [33].

MTurk has become increasingly popular among the scientific community, often used to perform user studies, natural language processing [33] and other studies that benefit from a large data set collection and where anonymity is not a concern. MTurk enables data collection in large scale specially overcoming two major difficulties in using the Internet for data collection: the recruitment of participants and the compensation. Higher paid tasks become more competitive and thus likely to generate quicker results.

The demographics of the MTurk in 2014 is composed of more than 500.000 individuals from 190 countries, dominated by workers from US followed by India, with less of a quarter working from other locations [34].

Before 2012, Amazon accepted worldwide worker applications, but concerns with the quality of workers, the labour law and money laundering have led to a stricter registering policy. Previously registered workers had been subject to account verification: full names, address, bank account, and social security number.

The choice of using MTurk as means of collecting data for this specific experiment is substantiated by two main conditions: the first concerns the amount of data necessary for the study to be reliable and conclusive, which we stipulated in a number around 90 for each of the 4 groups. Comparatively, collecting the same amount of data in an attendance setting would require a much greater deal of time and resources. The second reason is that the MTurk population is supposedly very diverse, which is in fact a requirement for the audience we want to target for this experiment.

Preventing duplicates is a concern since we don't have access to an individualized completion of the survey, and since workers get compensated for each completed task, it becomes necessary to have a discriminatory system for duplicates. Besides warning workers that they cannot repeat the survey, we implemented two methods for repetition verification:

- 1) Store MTurk workers ID's: we ask participants to fill their unique MTurk ID in the text field and store the ID in a txt file next to the estimate. In case a duplicate is found, the participant will be disqualified and will thus not receive any payment.

- 2) Install a cookie on the worker computer that blocks the access to the survey if the worker has already submitted a previous answer.
- 3) Verification mechanism that validates the MTurk ID introduced by the participant (in our database) against the current ID's database (on Amazon MTurk platform).

### 3.2.2. Reliability of data

We consider the data we gathered to be reliable. Due to the virtual nature of the participation in this experiment, we had no possibility of controlling the answers or the estimation process. Even though MTurk is widely used in academic research [34] we are aware that the estimates can be biased due to many factors we could not control, such as attention, joke, duplicates, errors, or simply users not understanding the interface or the task.

One central concern regarding the method of acquiring estimations was the fact that we had no control or access to how well participants understood the hints. The risk that participants might have ignored or not understood the information we provided next to the picture of the jellybeans jar is a fact that needs to be considered in the context of this study, results should be interpreted with a granular perspective. Additionally, we are also aware that the type of task requires a spatial understanding of a tri-dimensional object represented by a photograph.

We tried however to mitigate those limitations by implementing a reward prize and an optimized experience to avoid dismissive estimates. Nevertheless, this platform was extremely helpful in gathering an otherwise incomprehensible amount of data, and even facing inevitable uncertainty, the results are comparably consistent.

### 3.2.3. Erased data

Data was dynamically stored every time a participant made a guess. We collected the group number (to which participants were randomly assigned), the value of the estimation (a whole number no larger than 100 000) and the participant Mturk ID. We stored this information on a text file with the following format:

```
0,2114,A3T90ZWPBV0MCI
```

Additionally, to prevent duplicates, we kept a text file with all Mturk IDs and verified against this list every time a participant submitted their ID on the page.

Because the nature of this experiment does not allow us to physically control the estimation process we needed to adopt extra measures to maximize our certainty about honesty and attention. On the Mturk page, we had access to the Completion Time - one important variable presumably indicative of attention and/or

engagement of participants. We found that 10 seconds were the very minimum amount of time necessary to properly read the instructions, make the estimation and then paste the given code onto the MTurk survey page. Therefore, values inferior to 10 seconds were not considered and were erased (12 entries).

Another reason for excluding results was the apparent randomness of extremely high values when compared to the average, therefore we considered values above 100,000 to be the result of a careless estimate. Fewer cases were found with the incorrect completion code, which were also erased. A complete list of the erased data can be found in Appendix A on page 52.

#### **3.2.4. Considerations about erased files**

Although for the 'no information' group (group 0) erasing files made no difference, for all the information groups the erased values eventually explained variations in our data values since estimations occurred in close relation with others. Nonetheless we have considered these estimates to pollute the end results and decided to exclude them from our final data set.

## 4. Results

In this chapter we analyse the results of groups individually. We focus on the analysis of the central tendency of data and calculate the arithmetic mean, geometric mean, median, standard deviation, and plot the data distribution in percentiles for overview.

### 4.1. Individual group analysis

#### 4.1.1. Group 0: no influence

The 'no influence' group shows an even distribution of estimates. There is no correlation with previous estimates but values tend to consistently be below 2000 with few intermittent high estimates pushing the mean up to 3372.

The difference of the mean of all estimates to the real value (3150) is 222, 7% more. Comparable to the results of the limited experiments found in literature, a 7% difference to the real value denotes a good indicator of wisdom of the crowds.

FIGURE 8: DISTRIBUTION AND GLOBAL STATISTICS FOR GROUP 0

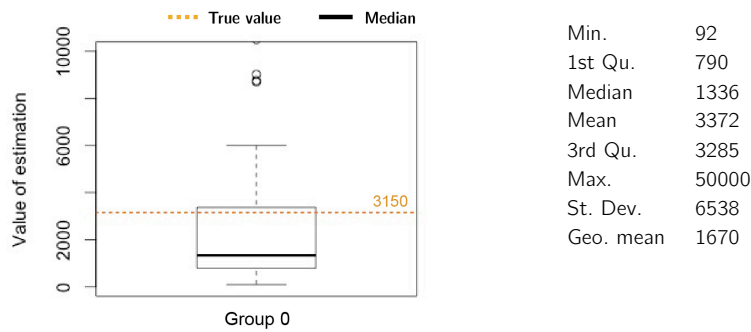


FIGURE 9: DISTRIBUTION OF ESTIMATES FOR GROUP 0 (N=98)

*No influence group: participants had no access to the estimates of others.*

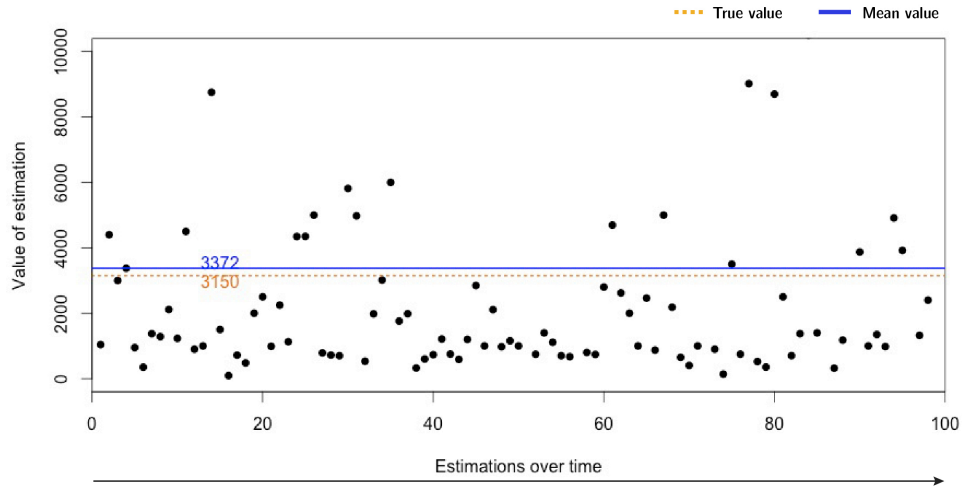


TABLE 5: LIST OF VALUES ORDERERED BY INPUT IN GROUP 0

*The list shows the values as introduced by participants over time (from left to right). The underlined values refer to values above 10000 not shown in the preceding figure.*

1040	4400	3000	3375	950	350	1374	1285	2114	1230	4500	900	999
8756	1500	92	718	480	2000	2500	988	2247	1128	4346	4350	5000
786	720	700	5816	4977	532	1980	3015	6000	1760	1984	325	600
732	1212	750	589	1200	2847	1000	2109	975	1156	1000	50000	746
1400	1111	700	672	12013	800	740	2800	4695	2620	2000	1000	2463
872	5000	2183	650	404	1000	12385	900	138	3500	748	9020	522
353	8700	2500	702	1375	10500	1400	<u>37142</u>	320	1178	<u>10957</u>	3872	1000
1350	981	4913	3920	<u>12150</u>	<u>1323</u>	2400						

#### 4.1.2. Group 1: the 5 best out of 10 random estimates

The distribution of estimates in group 1 shows a remarkable trend of estimates to follow an ascending path with little variation, then stabilizing around the true value by the 60<sup>th</sup> participant.

In this group we showed participants the five closest values to the true value out of ten random estimates. To the first participant no information was shown. To the 2nd participant it showed one guess (the previous), to the 3rd it showed the two previous guesses and so on until the 5th participant. When the number of guesses was more than 5 but less than 10, participants 6th to 11th were shown a random selection of 5 guesses from the previous guesses. When the number of guesses reached 10, it chose the 10 closest guesses and randomly showed 5 of those to the participant.

Figure 10 shows how the information method mechanics influenced the early guesses and how it progressed after the 60<sup>th</sup> guess into a steady gravitation around the true value (3150), which denotes a positive impact of social information even if partially randomized.

The average of guesses, using the arithmetic mean, is 3626, an increase of 486 when compared with the true value, 15% more.

FIGURE 10: DISTRIBUTION AND GLOBAL STATISTICS FOR GROUP 1

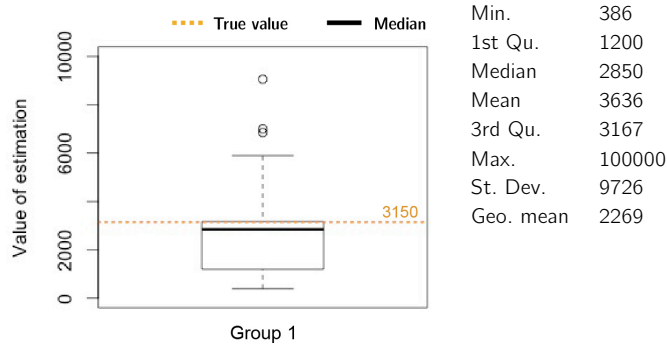


FIGURE 11: DISTRIBUTION OF ESTIMATES FOR GROUP 1 (N=105)

5 best out of random 10 group: participants had

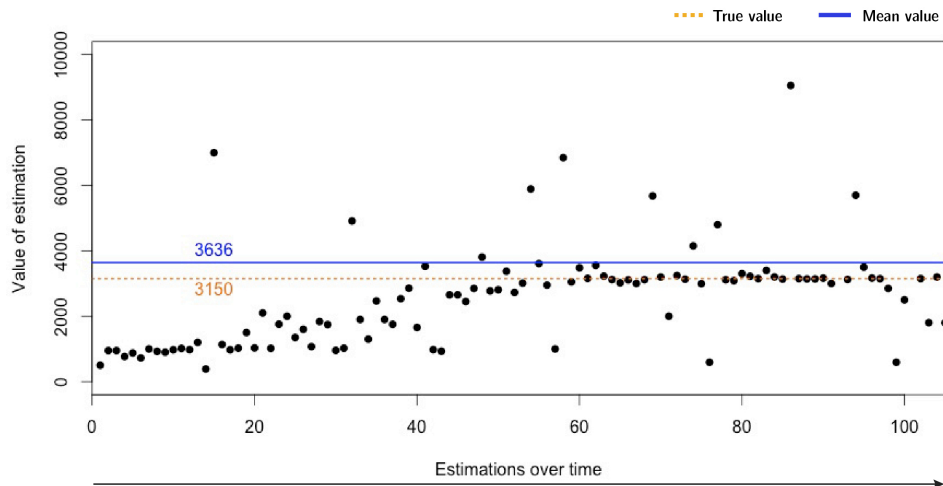


TABLE 6: LIST OF VALUES ORDERED BY INPUT IN GROUP 1

The list shows the values as introduced by participants over time (from left to right). The underlined values refer to values above 10000 not shown in the preceding figure.

500	950	950	767	875	724	1000	925	900	974	1017	978
1200	386	7000	1135	974	1025	1500	1031	2100	1019	1757	2000
1350	1600	1071	1836	1745	956	1023	4913	1900	1300	2467	1900
1753	2536	2857	1656	3526	978	930	2654	2654	2450	2850	3808
2774	2811	3375	2727	3012	5890	3610	2952	1000	6847	3050	3480
3158	3555	3225	3120	3020	3111	3000	3117	5678	3199	2000	3243
3132	4150	2993	592	4800	3116	3087	3302	3222	3150	3401	3201
3136	9054	3151	3140	3140	3165	3000	100000	3123	5700	3500	3167
3148	2850	592	2500	18000	3151	1803	3200	1800			



### 4.1.3. Group 2: bracketed mode

The bracketed mode group shows an even distribution of guesses with the lowest compared standard deviation of 2416. The arithmetic mean is 3485, a 10,6% above the true value (3150). Participants in this group were exposed to the most common intervals of previous estimations which results in a very clear contained pattern with very few deviations - the group with the lower standard deviation (2420).

FIGURE 12: DISTRIBUTION AND GLOBAL STATISTICS FOR GROUP 2

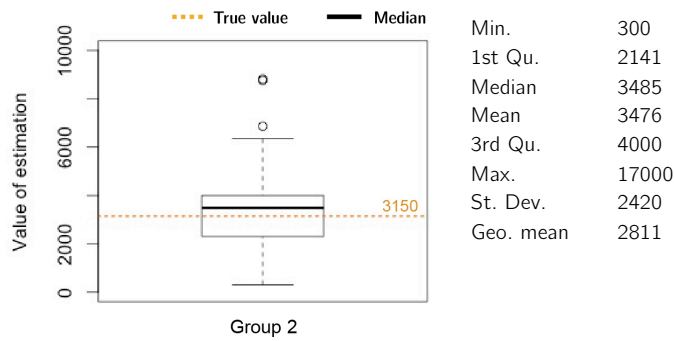


FIGURE 13: DISTRIBUTION OF ESTIMATES FOR GROUP 2 (N=86)

*Bracketed mode*

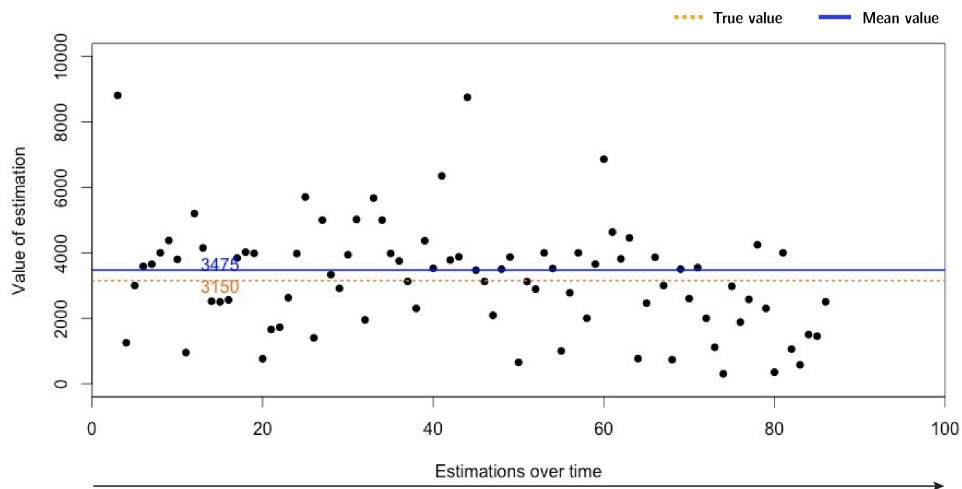


TABLE 7: LIST OF VALUES ORDERED BY INPUT IN GROUP 2

*The list shows the values as introduced by participants over time (from left to right). The underlined values refer to values above 10000 not shown in the preceding figure.*

17000	12500	8813	1250	3000	3586	3651	4001	4376	3800	950	5200	4150	2520
2500	2560	3840	4021	3983	760	1657	1724	2626	3976	5706	1400	5000	3333
2912	3942	5021	1950	5675	5000	3980	3752	3126	2300	4368	3525	6352	3780
3878	8755	3469	3124	2088	3500	3871	650	3120	2890	4000	3520	1000	2777
3999	2701	3655	6862	4633	3816	4456	765	2462	3865	3000	733	3500	2600
3550	1998	1111	300	2976	1880	2575	4246	2300	350	4000	1055	575	1500
1450	2501												

#### 4.1.4. Group 3: full information

The distribution of estimates in group 3 indicates no particular pattern with regards to the influence of estimates. As participants guessed the number of jellybeans, they had full access to all the previous estimates in the form of a dispersion map. By presenting participants with a visual map of previous guesses, we let participants generate their own personal *aggregation* of previous guesses and make a new guess based on that information and their own heuristics, instead of presenting them with a ready-made aggregation as in [25].

The arithmetic mean of all estimates is 3387, 7,52% above the true value (3150). The standard deviation is large but it comes quite close to the results of group 0 where no information is disclosed.

FIGURE 14: GLOBAL STATISTICS AND DISTRIBUTION FOR GROUP 3

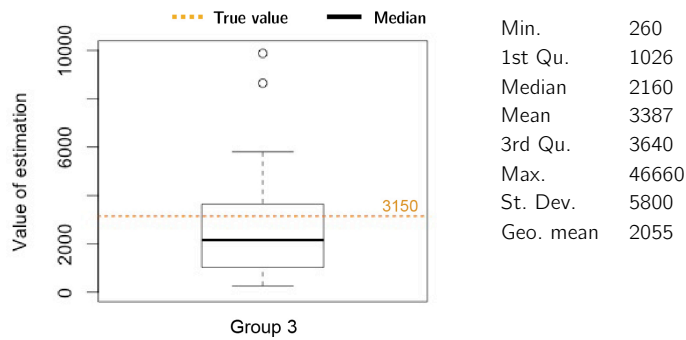


FIGURE 15: DISTRIBUTION OF ESTIMATES FOR GROUP 3 (N=91)

*Full information*

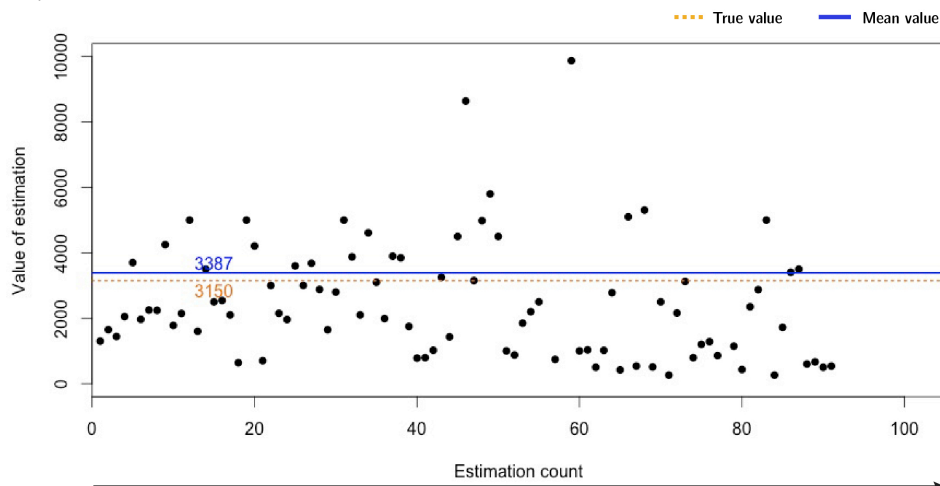


TABLE 8: LIST OF VALUES ORDERERED BY INPUT IN GROUP 3

*The list shows the values as introduced by participants over time (from left to right). The underlined values refer to values above 10000 not shown in the preceding figure.*

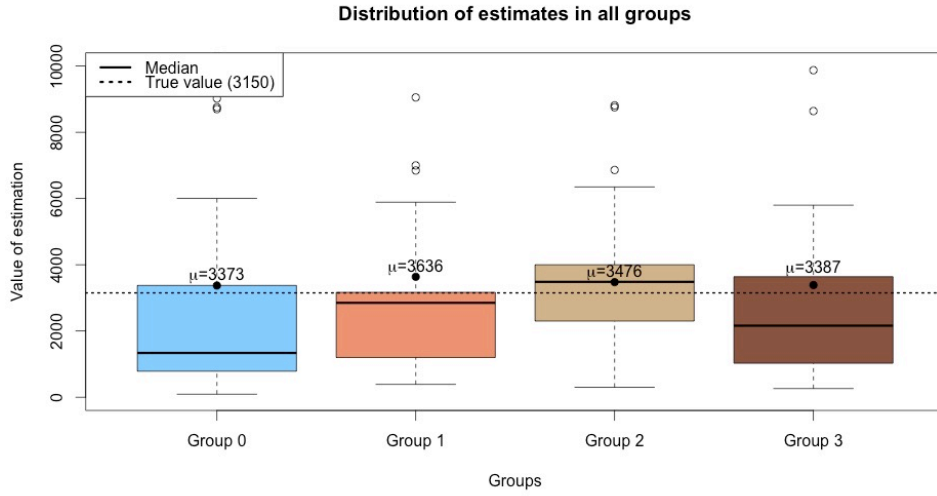
1300	1650	1440	2050	3700	1966	2250	2240	4250	1780	2146	5000	1600	3500
2500	2543	2100	640	5000	4208	700	3000	2150	1960	3600	3002	3680	2881
1647	2800	5000	3876	2100	4612	3100	1990	3897	3847	1750	780	793	1020
3250	1429	4500	8641	3152	4983	5800	4500	1000	875	1850	2200	2500	12621
742	30254	9875	1000	1032	500	1017	2780	420	5100	537	5306	510	2500
260	2160	3125	793	1200	1284	857	46656	1145	431	2350	2875	5000	260
1720	3405	3500	600	666	500	536							

## 4.2. Group performance analysis

### 4.2.1. Arithmetic mean

A Shapiro-Wilk test ( $p > 0.05$ )<sup>2</sup> showed that the sample data, when logarithmized, was not approximately distributed. Therefore, the arithmetic mean is a better predictor of WoC than the geometric mean used in [2] and it is considered as the primary indicator of the wisdom of the crowd effect in our case. Group 0 comes closer to the true value by 7% immediately followed by Group 3 with 7,52%. Group 1 and 2 account for 15% and 10,6% variation respectively.

FIGURE 16: GROUP DISTRIBUTION OVERVIEW



A Mann-Whitney test ( $U=3864$ ,  $Z=-1.5821$ ,  $p=0.057$ ) indicated that the difference between the medians, hence the distribution location of the control group (group 0) and the full information group (group 3) is not significantly different. The differences of the medians of group 1 and 2 showed however a significant difference from the control group<sup>3</sup>. This led us to conclude that the accuracy of estimates does not change significantly if a group is exposed to social influence in its full form.

### 4.2.2. Measures from Diversity Prediction Theorem: Individual error

The progression of Average Individual Error (AE) over time calculates the absolute value of the difference of every estimate to the truth value  $\sqrt{(truth - \hat{x})^2}$ . Figure 17 displays the evolution of AE, which is particularly relevant to observe for the social influence groups, since each estimate is correlated with the aggregation method used to provide social information about the previous estimates. Groups 1

<sup>2</sup> The Shapiro-Wilk test returned the same p value of  $2.2e-16$  for groups 0, 1 and 3, and  $1.726e-10$  for Group 2, denoting that sample data for all groups is not normally distributed. See appendix C on page 58 for full test details.

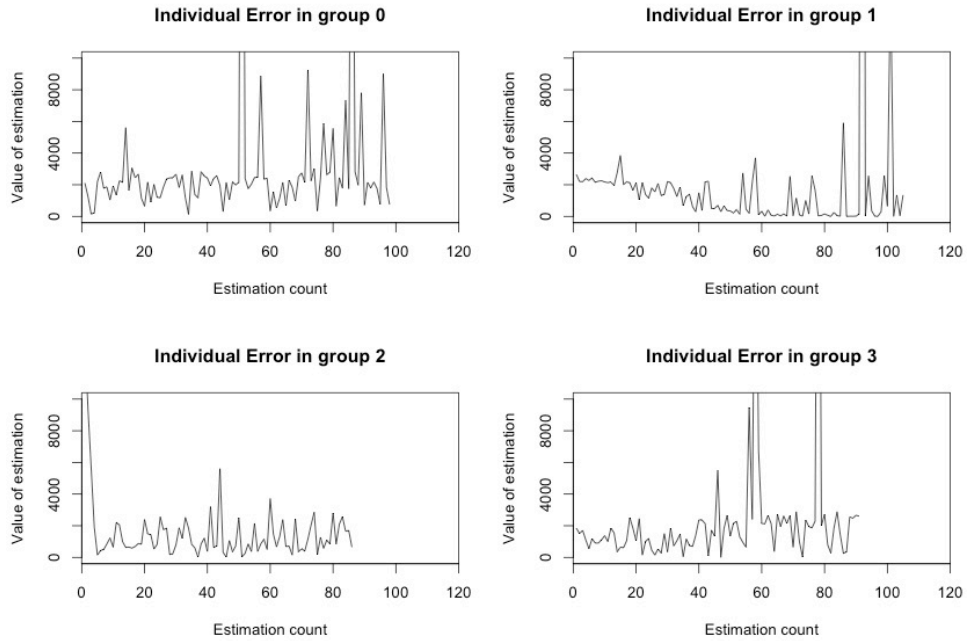
<sup>3</sup> See Appendix D: Mann-Whitney U-Test on page 54 for detailed results.

and 2 (5 out of 10 random and bracketed mode) where the aggregated methods provided less information, indicate a stronger correlation with previous estimates exhibiting a visible pattern in AE. Group 1 shows a declining trend in AE that stabilizes around the 60<sup>th</sup> estimate with a very low error. This demonstrates that the hint impacts the estimates significantly and that the strength of this effect can be tracked back to the qualitative aspect of the hint. So by showing a hint that entails the best estimates in relation to the truth (even if randomized) a low AE trend was established, although in our case alternated with some extreme high errors close to the 90<sup>th</sup> estimate.

Group 2 shows a smaller variance in AE compared to the other groups. Errors are systematically below 4000 and the hint seems to prevent extreme high estimates as seen in all the other groups. Group 0 and 3 display a similar pattern but group 0 entails a higher AE.

FIGURE 17: INDIVIDUAL ERROR OVER TIME (SEQUENCE OF ESTIMATES)

*Progression of individual error over time measures the absolute distance from each estimation to the true value.*



#### 4.2.3. Measures from Diversity Prediction Theorem: Diversity and Collective Error

The diversity is the measurement of estimates in terms of its variance from the mean, denoted as the average of the squared difference of estimates to the mean of estimates  $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ , where  $\bar{x}$  is the sample mean average of all estimates and  $n$  is the sample size. We use the unbiased sample variance ( $n - 1$ ) which produces a more accurate estimation of the true variance of a sample with regards to the population variance.

The collective error, as seen on the Diversity Prediction Theorem on page 9, equals the average individual error minus diversity, and it is the squared deviation of the group’s mean from the truth  $(truth - \bar{x})^2$ , also known as population bias [2]. Our results show that independent group (group 0) holds more diversity than the social treatment groups, and consequently the collective error is also the lowest.

TABLE 9: DIVERSITY IN ALL GROUPS

*Diversity calculated as the sample variance of estimate.*

	Group 0	Group 1	Group 2	Group 3
Diversity	42,316,034	93,710,910	5,789,851	33,281,044
Average error	42,365,540	93,947,477	5,896,014	33,337,359
Collective error	49,506	236,566	106,162	56,315

However, to assess whether both groups' diversity differed significantly, we performed a Levene’s statistical variance test, which measured the difference in the dispersion of estimates between the independent group (Group 0) and the social groups (1, 2 and 3). The test compared the equality of variance<sup>4</sup> for non-normally distributed data by converting each value by calculating the absolute deviation of observations from the median. The tests failed to reject the null hypothesis ( $p > 0.05$ )<sup>5</sup> for all groups, there is insufficient evidence to claim that the variances are not equal, which in our case concludes that both groups present an approximate variance of the distribution of estimates.

#### 4.2.4. Dispersion of estimates: Wisdom of the Crowd Indicator

According to [2], we can also consider the WoC effect to exist when the aggregate is close to the truth relative to the dispersion of the sample. Then, a sample with a small collective error but larger dispersion shows more WoC than a sample with the same small collective error but small dispersion – there is no crowd wisdom if the crowd does not outperform the individuals by a considerable magnitude.

We can illustrate this with the example of a government that needs an advice or a prediction. If the predictions are spread narrowly around the wrong value, a decision maker would gain confidence in information that is incorrect. The

<sup>4</sup> Variance is a spread measure that represents the average squared deviations of the estimates to the mean.

<sup>5</sup> Levene’s test returned a  $p$  value  $> 0.05$  for homogeneity of variance between group 0 and all other groups, denoting that the variance of the estimates is not significantly different across all groups. See appendix B on page 57 for detailed results.

clustering around the wrong value makes the group less wise, as the truth is not located centrally but in outer regions of the range of estimates.

This concept of “bracketing the truth” is illustrated in For our data set, we have adopted the WoC indicator applied in [2] to visualise the range of estimates in relation to the median as  $\{i | \hat{x}_i \leq \text{truth} \leq \hat{x}_{n-i+1}\}$ . In table 10 we can see the estimates sorted in ascending order and in grey the range of estimations in relationship to the true value. The diamond shape marks the centre of the ordered estimates while the dark highlights indicate the two values between the truth-value. The WoC indicator is measured in the number of steps or values between the centre of estimates and the truth.

Table 10, Appendix E on page 58. However, this model to quantify the WoC has shown to be inadequate for our data<sup>6</sup>. Firstly, because our groups have different sizes, so a measure that is based on the median will evidently not provide a reliable comparison between groups. Secondly, our data is not normally distributed so the median is not a measure that we can use to quantify the bracketing of the truth.

---

<sup>6</sup> The wisdom of the crowd as defined in [2] is a measure to quantify the WoC effect for normally distributed data, and therefore it uses the median to quantify the bracketing of the truth (clustering around the truth value with respect to median). In [2] the median is defined as the appropriate measure for aggregation and it coincides with the geometric mean for log-normal distribution in their case, but it is not a generalised measure for all distribution types.

## 5. Conclusions

### 5.1. Theoretical implications

The results in our study suggest that social influence might not inhibit the WoC as inferred in [2]. Quite the opposite, we verified that providing unaggregated social information equally produced an accurate estimation of the right number of jellybeans.

The amount of information carried in the hints we provided is relative to how aggregated or compressed the collective information was shown to each subject. More aggregated information, as displayed group 1 (five best estimates out of ten random), carries less original information about the estimates, therefore more compressed information was observed as limiting the effect of WoC. Group 3 (which displays full information) performs almost as well as group 0 (which displays no information) with a slight difference of 0,52%. Groups 1 and 2 (15% and 10,6% respectively) can also prudently be considered successful when compared with other studies [2] where differences to the truth was at its best 11%.

The Levene's test to assess the homogeneity of variances also indicated that there is no significant difference in the diversity (variance of estimates) of all groups, denoting that social information does not alter the essence of diversity of estimates.

When comparing the means (Mann-Whitney U-test) of social treatment groups against the control group, we have seen that the distribution of the estimates regarding its centre (the median) is not significantly different between the control group and the full information group, but different between the control group and groups 1 and 2, implying that distribution of estimates is similar between group 0 and group 3.

In the context of the Diversity Prediction Theorem, group 0 has the lowest Collective Error score immediately followed by group 3. Although group 0 holds more diversity<sup>7</sup> and has a slightly better mean<sup>8</sup> than group 3, the variance difference between the groups is not significant<sup>9</sup>. So, considering the mean distance to the truth (which is 7% and 7,53% respectively) and the homogeneity of variances of both groups, we can easily conclude that the difference between the independent group and social full information groups is not significant.

---

<sup>7</sup> As seen in Table 9: Diversity in all groups, page 45.

<sup>8</sup> As seen in Figure 16: Group distribution overview, page 43.

<sup>9</sup> See Appendix B: Levene's test for detailed test results, page 54



This confirms, to a certain extent and limited to the scope of this study, our hypothesis that social influence may be consistent with the WoC effect and not inhibit as suggested in [2][3][4]. In particular, we have observed that the key for a more efficient use of social influence resides in the form social information is presented.

By displaying full information we enable multiple comparisons between options: several iterations between individual heuristics and the information of others, where the risk of copying cascades of poor estimates is unlikely because it would take several individuals to come to the same conclusion independently to reach the same quorum threshold [11].

In the actual context, this study contributes to a still young discussion about the impact of social interaction on the WoC effect. There are still few experiments exploring this phenomenon and standards have not yet emerged. When compared to [25], we have introduced a more realistic concept to present social information (Group 3) based on a full disclosure of information, which we believe to be a more realistic real world approach to mimic social imitation processes rather than showing an aggregation of the best estimates so far. Particularly when compared to [2], not only does our study refute the hypothesis that social influence undermines per se the wisdom of the crowd effect, but we also offer substantial evidence to support that the form social influence occurs is determinant to attain the wisdom of the crowd effect.

We believe that the experiment we designed helps to shed some light onto the enabling conditions for wisdom of crowd effect to occur and establishes a new ground for discussion regarding the degrees of social information and the impact it has in decreasing a crowd's diversity.

## 5.2. Future research

The investigation we undertook only scratched the surface of the immense potential of the WoC. Many aspects of its mechanics need further research to reach a solid standard for applicability: the types of problems, the aggregation methods and the shapes social influence can take are just the very initial aspects harnessing the wisdom of the crowds.

We have specifically looked at the impact of social influence considering three different information degrees concluding that providing non-aggregated full information performs nearly as well as providing no information. In the sequence of our study, further work is necessary to assess the strength of our results with respect to information degrees of social influence.

The aggregation formula applied to the social influence mechanism - in our case in the form of a hint - has shown to have impacted the WoC effect. In the social influence groups, we have seen that disclosing full information produces better results than the other two types of aggregation tested, denoting that conveying more information, hence less aggregated hints, is better to attain a more accurate WoC effect.

However, multiple aggregate formulas are possible. In future research, it would be of interest to understand the impact of aggregates that relates to the true value. In our experiment, group 2 partially included this qualifying aspect by presenting a hint with the five best estimates out of ten random. The group performed poorly compared to the other groups. One reason could be that the qualitative aspect of the hint prevented participants from iterating actively with the hint information and just accepting the best as an absolute truth-value.

Further investigation regarding the variations of information degrees with regards to the truth would be beneficial to enhance our understanding of the underlying mechanisms of individual decision-making and how it impacts collective performance.

## 6. References

- [1] F. Galton, “Vox Populi”, *Nature*, vol. 75, pp. 450–451, 1907.
- [2] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, “How social influence can undermine the wisdom of crowd effect”, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 22, pp. 9020–5, May 2011.
- [3] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental Study of Inequality and Cultural Market”, vol. 311, no. February, 2006.
- [4] S. Krause, R. James, J. J. Faria, G. D. Ruxton, and J. Krause, “Swarm intelligence in humans: diversity can trump ability”, *Anim. Behav.*, vol. 81, no. 5, pp. 941–948, May 2011.
- [5] D. J. T. Sumpter and S. C. Pratt, “Quorum responses and consensus decision making”, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 364, no. 1518, pp. 743–53, Mar. 2009.
- [6] T. W. Malone, R. Laubacher, and C. Dellarocas, “Harnessing Crowds : Mapping the Genome of Collective Intelligence”, 2010.
- [7] R. L. Goldstone and T. M. Gureckis, “Collective Behavior,” *Top. Cogn. Sci.*, vol. 1, no. 3, pp. 412–438, Jul. 2009.
- [8] “Handbook Collective Behavior”, *Intelligence, MIT Center for Collective*, 2012.  
[Online]. Available: [http://scripts.mit.edu/~cci/HCI/index.php?title=Main\\_Page](http://scripts.mit.edu/~cci/HCI/index.php?title=Main_Page).  
[Accessed: 22-Jun-2016].
- [9] M. Ziewitz, “Can Crowd Wisdom Solve Regulatory Problems ? Can crowd wisdom solve regulatory problems ? A review and some provocations”, in *1st Berlin Symposium on Internet and Society*, 2011.
- [10] S. C. Pratt, “Collective Intelligence”, vol. 1, pp. 303–309, 2010.
- [11] J. Krause, G. D. Ruxton, and S. Krause, “Swarm intelligence in animals and humans”, *Trends Ecol. Evol.*, vol. 25, no. 1, pp. 28–34, Jan. 2010.
- [12] B. Bahrami, K. Olsen, D. Bang, A. Roepstorff, G. Rees, and C. Frith, “What failure in collective decision-making tells us about metacognition”, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 367, no. 1594, pp. 1350–65, May 2012.
- [13] S. Page, *Difference, The*. Princeton University Press, 2007.
- [14] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry, “The wisdom of the crowd in combinatorial problems”, *Cogn. Sci.*, vol. 36, no. 3, pp. 452–70, Apr. 2012.
- [15] C. J. Tessone and F. Schweitzer, “Effects of Social Influence on the Wisdom of Crowds”, 2012.
- [16] M. Steyvers, M. Lee, B. Miller, and P. Hemmer, “The Wisdom of Crowds in the Recollection of Order Information”, 2009.
- [17] J. M. Leimeister, “Collective Intelligence”, *Bus. Inf. Syst. Eng.*, vol. 2, no. 4, pp. 245–248, Jun. 2010.
- [18] R. S. Burt, “Structural Holes and Good Ideas”, *Am. J. Sociol.*, vol. 110, no. 2, pp. 349–399, 2004.

- [19] R. P. Larrick, A. E. Mannes, and J. B. Soll, "The social psychology of the wisdom of crowds", *Front. Soc. Psychol. Soc. Judgm. Decis. Mak.*, pp. 227–242, 2012.
- [20] C. Wagner and T. Vinaimont, "Evaluating the wisdom of crowds", vol. XI, no. 1, 2010.
- [21] M. D. Lee, S. Zhang, and J. Shi, "The wisdom of the crowd playing The Price Is Right", *Mem. Cognit.*, vol. 39, no. 5, pp. 914–23, Jul. 2011.
- [22] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, "Evidence for a collective intelligence factor in the performance of human groups", *Science*, vol. 330, no. 6004, pp. 686–8, Oct. 2010.
- [23] J. E. McGrath, "Groups: Interaction and Performance", Prentice-Hall, Inc., 1983.
- [24] J. Surowiecki, "The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations", Abacus, 2004.
- [25] A. J. King, L. Cheng, S. D. Starke, and J. P. Myatt, "Is the true 'wisdom of the crowd' to copy successful individuals?", *Biol. Lett.*, vol. 8, no. 2, pp. 197–200, Apr. 2012.
- [26] B. J. Miller and M. Steyvers, "The Wisdom of Crowds with Communication", pp. 1292–1297, 2009.
- [27] H. Rauhut and J. Lorenz, "The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions", *J. Math. Psychol.*, vol. 55, no. 2, pp. 191–197, Apr. 2011.
- [28] E. Vul and H. Pashler, "Measuring the crowd within: probabilistic representations within individuals", *Psychol. Sci.*, vol. 19, no. 7, pp. 645–7, Jul. 2008.
- [29] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers", *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 46, pp. 16385–9, Nov. 2004.
- [30] I. D. Couzin, "Collective cognition in animal groups", *Trends Cogn. Sci.*, vol. 13, no. 1, pp. 36–43, Jan. 2009.
- [31] T. Sasaki and S. C. Pratt, "Emergence of group rationality from irrational individuals", *Behav. Ecol.*, vol. 22, no. 2, pp. 276–281, Jan. 2011.
- [32] R. L. Goldstone, a. Jones, and M. E. Roberts, "Group path formation", *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 36, no. 3, pp. 611–620, May 2006.
- [33] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson, "Who are the Turkers ? Worker Demographics in Amazon Mechanical Turk", in *ACM CHI Conference in 2010*, 2010, pp. 1–5.
- [34] C. Paolacci, "Understanding Mechanical Turk," 2014.
- [35] "NIST/SEMATECH e-Handbook of Statistical Methods," 2013. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm>. [Accessed: 11-Jun-2016].
- [36] A. Hart, "Mann-Whitney test is not just a test of medians : differences in spread can be important," *BMJ Br. Med. Journal*, 323(7309), pp. 391–393, 2001.
- [36] A Facts about Google and Competition. [Online]. Available: <http://web.archive.org/web/20111104131332/http://www.google.com/competition/howgooglesearchworks.html> [Accessed: 11-Jun-2016]

## 7. Appendices

### Appendix A: List of erased values

#### VALUES ERASED DUE TO EXTREME VALUES

##### Group 0

0,225256,AZA4W311KW59S

##### Group 1

Position 1 1,100000,A2ECHY8E6SX7KP

##### Group 3

Position 41 3,1630271267,A2XQW7M1267TX  
Position 64 3,6720000,A1AFFVCA003FON  
Position 83 3,1304250,A3D0S6TR16HHZW  
Position 97 3,482530223, A2RCPY5Y131CXE  
Position 100 3,105000,A2SC0KSFYNW5IR

#### VALUES WITH COMPLETION TIME < 10"

Position 56, 2,2794,ADSSLREQARFS0  
Position 44, 2,3456,A3HEFMJ50IMTY6  
Position 65,1,3020,A21NRMZFK7QCJW  
Position 15,0,4000,A3UEFIZ8PF8281  
Position 59, 2,2000,A5CHEN7F50C03  
Position 28, 0,5500,A2X6K5T4P6GXTY  
Position 6, 0,350,A2DNSD743W40C2  
Position 42, 3,753,A10RNK847NK97J  
position 22, 3,2380,A1ENHFQ5X0XG6I  
position 31, 3,1200,A1NBMA287PWN0T  
position 45, 1,2654,A3A0J29Z72NSC3  
position 40, 3,1215,A3IPMSDYZPFVIL

## Appendix B: Levene's test

The Levene's test is a non-parametric test to determine homogeneity of variances between the groups 0 and 3. The test is used to test the assumption that both groups have equal variance. Variance is the square root deviation of a variable from its mean, measuring how disperse the data is from its mean. However, because our data follows a skewed distribution, we use the extended version of the Levene's test introduced by Brown and Forsythe (1974), which uses the median as reference instead of the mean as it has been proven to provide more robust results statistics [35]. The test returns the degrees of freedom (Df), the F-value and the P-value ( $\Pr(>F)$ ). To validate the null hypothesis only the P-value is considered for a significance level at 0.05.

The Levene's test is defined as follows:

### Group 0 - 1

$H_0: \sigma^2 \text{ Group 0} = \sigma^2 \text{ Group 3}$

The null hypothesis is that both groups have the same variance ( $p \geq 0.05$ ).

$H_a: \sigma^2 \text{ Group 0} \neq \sigma^2 \text{ Group 3}$

The alternative hypothesis is that the variance is significantly different ( $p \leq 0.05$ ).

**Significance level:** 0.05

**Result:**  $p = 0.7204$

It fails to reject  $H_0$  ( $p > 0.05$ ), therefore **groups are significantly homogenous**.

```
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 1 0.1284 0.7204
201
```

### Group 0 - 2

$H_0: \sigma^2 \text{ Group 0} = \sigma^2 \text{ Group 2}$

The null hypothesis is that both groups have the same variance ( $p \geq 0.05$ ).

$H_a: \sigma^2 \text{ Group 0} \neq \sigma^2 \text{ Group 2}$

The alternative hypothesis is that the variance is significantly different ( $p \leq 0.05$ ).

**Significance level:** 0.05

**Result:**  $p = 0.1269$

It fails to reject  $H_0$  ( $p > 0.05$ ), therefore **groups are significantly homogenous**.

```
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 1 2.3516 0.1269
182
```

### Group 0 - 3

$H_0: \sigma^2 \text{ Group 0} = \sigma^2 \text{ Group 3}$

The null hypothesis is that both groups have the same variance ( $p \geq 0.05$ ).

$H_a: \sigma^2 \text{ Group 0} \neq \sigma^2 \text{ Group 3}$

The alternative hypothesis is that the variance is significantly different ( $p \leq 0.05$ ).

**Significance level:** 0.05

**Result:**  $p = 0.6735$

It fails to reject  $H_0$  ( $p > 0.05$ ), therefore **groups are significantly homogenous**.

```
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 1 0.1781 0.6735
187
```

### Appendix C: Shapiro-Wilk test for normality

The Shapiro-Wilk test was used to assess the normality of the logarithmized data distribution in both. The Shapiro-Wilk test is defined as follows:

$H_0$ : Data is normally distributed ( $p \geq 0.05$ )

$H_a$ : Data is not normally distributed ( $p \leq 0.05$ )

**Significance level:** 0.05

#### Results:

$p$  (Group 0) = 2.2e-16

$p$  (Group 1) = 2.2e-16

$p$  (Group 2) = 1.726e-10

$p$  (Group 3) = 2.2e-16

It rejects the  $H_0$  ( $p < 0.05$ ), therefore all groups are not normally distributed.

```
Shapiro-Wilk normality test

> shapiro.test(Group0$estimate)
data:  Group0$estimate
W = 0.4204, p-value < 2.2e-16

> shapiro.test(Group1$estimate)
data:  Group1$estimate
W = 0.1825, p-value < 2.2e-16

> shapiro.test(Group2_100$estimate)
data:  Group2_100$estimate
W = 0.7619, p-value = 1.726e-10

> shapiro.test(Group3$estimate)
data:  Group3$estimate
W = 0.3945, p-value < 2.2e-16
```

## Appendix D: Mann-Whitney U-Tests

The Mann-Whitney test is a non-parametric test that allows for a comparison between a treatment or a condition in two groups without assuming that the data is normally distributed. It is the equivalent of a t-test for non-normally distributed data that detects the differences in shape and spread as well as the differences between the medians [36]. Because it uses the median it is not as sensitive to outliers as the t-test, which it is based on differences between the means. Although a normality test (Shapiro-Wilk test  $p < 0.05$ ) has indicated that our data for all groups is not normally distributed, the Mann-Whitney considered our data to be approximately normal and therefore the Z-values could be used. Because the difference between p values of the one tail and the two tail tests delivered was very little, we opted to consider the two tailed test results valid for our data.

The Mann-Whitney U-test is defined as follows:

**H<sub>0</sub>:** The null hypothesis asserts that the medians of the two samples are identical; therefore there is no significant difference between the two samples ( $p \geq 0.05$ ).

**H<sub>a</sub>:** The alternative hypothesis asserts that the medians of the two samples are not identical ( $p \leq 0.05$ ).

**Significance level:** 0.05

CONTROL GROUP (GROUP 0) - FULL INFORMATION GROUP (GROUP 1)

### Results:

(One tailed test:  $U = 3970$ ;  $Z = 2.8082$ ;  $p = 0.00248$ )

**Two tailed test:**  $U = 3970$ ;  $Z = 2.8082$ ;  **$p = 0.00496$**

The U-value is 3970. The distribution is approximately normal. The result is significant at  $p \leq 0.05$ . Therefore, the Z-value can be used.

It rejects H<sub>0</sub> ( $p \leq 0.05$ ), therefore both groups are **significantly different**.

### Result Details (Two Tailed test)

<b>Group 0</b> Sum of ranks: 8821 Mean of ranks: 90.01 Expected sum of ranks: 9996 Expected mean of ranks: 102 U-value: 6320 Expected U-value: 5145	<b>Group 1</b> Sum of ranks: 11885 Mean of ranks: 113.19 Expected sum of ranks: 10710 Expected mean of ranks: 102 U-value: 3970 Expected U-value: 5145
---	--

**Group 0 and Group 1 Combined**  
Sum of ranks: 20706  
Mean of ranks: 102  
Standard Deviation: 418.2463

### U and P Values

**By Meta Numerics**  
U-value: 3984  
P-value (left probability): 0.0028  
P-value (right probability): 0.9972

**By ALGLIB**  
P-value (combined): 0.0048



CONTROL GROUP (GROUP 0) - FULL INFORMATION GROUP (GROUP 2)

**Results:**

(one tail test:  $U = 2690$ ;  $Z = -4.2265$ ;  $p = 0$ )

**Two tail test:**  $U = 2690$ ;  $Z = -4.2265$ ;  **$p = 0$**

The Z-Score is -4.2265. The p-value is 0. The result is **significant** at  $p \leq 0.05$ . The U-value is 2690. The distribution is approximately normal. Therefore, the Z-value above can be used.

It rejects  $H_0$  ( $p \leq 0.05$ ), therefore both groups are **significantly different**.

**Result Details (Two Tailed test)**

<b>Group 0</b>	<b>Group 2</b>
Sum of ranks: 7541	Sum of ranks: 9479
Mean of ranks: 76.95	Mean of ranks: 110.22
Expected sum of ranks: 9065	Expected sum of ranks: 7955
Expected mean of ranks: 92.5	Expected mean of ranks: 92.5
U-value: 5738	U-value: 2690
Expected U-value: 4214	Expected U-value: 4214

<b>Group 0 and Group 2 Combined</b>
Sum of ranks: 17020
Mean of ranks: 92.5
Standard Deviation: 360.4604

**U and P Values**

By Meta Numerics  
U-value: 2702  
P-value (left probability): 0  
P-value (right probability): 1

By ALGLIB  
P-value (combined): 0.0001

CONTROL GROUP (GROUP 0) - FULL INFORMATION GROUP (GROUP 3)

### Results

(one tail test:  $U = 3864$ ;  $Z = -1.582$ ;  $p = 0.05705$ )

**Two tail test:**  $U = 3864$ ;  $Z = -1.582$ ;  $p = 0.1141$

The Z-Score is -1.5821. The p-value is 0.1141. The result is **not significant** at  $p \leq 0.05$ .

The U-value is 3864. The distribution is approximately normal. Therefore, the Z-value above can be used.

It **fails to reject**  $H_0$  ( $p \leq 0.05$ ), therefore both groups are **not significantly different**.

### Result Details (Two Tailed test)

Group 0	Group 3
Sum of ranks: 8715	Sum of ranks: 9240
Mean of ranks: 88.93	Mean of ranks: 101.54
Expected sum of ranks: 9310	Expected sum of ranks: 8645
Expected mean of ranks: 95	Expected mean of ranks: 95
U-value: 5054	U-value: 3864
Expected U-value: 4459	Expected U-value: 4459

Group 1 and Group 3 Combined
Sum of ranks: 17955
Mean of ranks: 95
Standard Deviation: 375.7681

### U and P Values

By Meta Numerics  
U-value: 3984  
P-value (left probability): 0.0028  
P-value (right probability): 0.9972

By ALGLIB  
P-value (combined): 0.0048

## Appendix E: Wisdom of the crowd indicator

For our data set, we have adopted the WoC indicator applied in [2] to visualise the range of estimates in relation to the median as  $\{i|\hat{x}_i \leq \text{truth} \leq \hat{x}_{n-i+1}\}$ . In table 10 we can see the estimates sorted in ascending order and in grey the range of estimations in relationship to the true value. The diamond shape marks the centre of the ordered estimates while the dark highlights indicate the two values between the truth-value. The WoC indicator is measured in the number of steps or values between the centre of estimates and the truth.

TABLE 10: WISDOM OF THE CROWD INDICATOR

List of estimates sorted in ascendant order.

Key													
◆ Center of estimates (median)													
■ Range of estimates in relation to the truth													
■ Truth interval between two values													
(Group 0, N=98, WoC Ind=25)													
92	138	320	325	350	353	404	480	522	532	589	600	650	672
700	700	702	718	720	732	740	746	748	750	786	800	872	900
900	950	975	981	988	999	1000	1000	1000	1000	1000	1040	1111	1128
1156	1178	1200	1212	1230	1285	1323	◆1350	1374	1375	1400	1400	1500	1760
1980	1984	2000	2000	2109	2114	2183	2247	2400	2463	2500	2500	2620	2800
2847	3000	3015	3375	3500	3872	3920	4346	4350	4400	4500	4695	4913	4977
5000	5000	5816	6000	8700	8756	9020	10500	10957	12013	12150	12385	37142	50000
(Group 1, N=105, WoC Ind=34)													
386	500	592	592	724	767	875	900	925	930	950	950		
956	974	974	978	978	1000	1000	1017	1019	1023	1025	1031		
1071	1135	1200	1300	1350	1500	1600	1656	1745	1753	1757	1800		
1803	1836	1900	1900	2000	2000	2100	2450	2467	2500	2536	2654		
2654	2727	2774	◆2811◆	2850	2850	2857	2952	2993	3000	3000	3012		
3020	3050	3087	3111	3116	3117	3120	3123	3132	3136	3140	3140		
3148	3150	3151	3151	3158	3165	3167	3199	3200	3201	3222	3225		
3243	3302	3375	3401	3480	3500	3526	3555	3610	3808	4150	4800		
4913	5678	5700	5890	6847	7000	9054	18000	100000					
(Group 2, N=86, WoC Ind=41)													
300	350	575	650	733	760	765	950	1000	1055	1111	1250	1400	1450
1500	1657	1724	1880	1950	1998	2088	2300	2300	2462	2500	2501	2520	2560
2575	2600	2626	2701	2777	2890	2912	2976	3000	3000	3120	3124	3126	3333
3469	◆3500	3500	3520	3525	3550	3586	3651	3655	3752	3780	3800	3816	3840
3865	3871	3878	3942	3976	3980	3983	3999	4000	4000	4001	4021	4150	4246
4368	4376	4456	4633	5000	5000	5021	5200	5675	5706	6352	6862	8755	8813
12500	17000												
(Group 3, N=91, WoC Ind=29)													
260	260	420	431	500	500	510	536	537	600	640	666	700	742
780	793	793	857	875	1000	1000	1017	1020	1032	1145	1200	1284	1300
1429	1440	1600	1647	1650	1720	1750	1780	1850	1960	1966	1990	2050	2100
2100	2146	2150	◆2160◆	2200	2240	2250	2350	2500	2500	2500	2543	2780	2800
2875	2881	3000	3002	3100	3125	3152	3250	3405	3500	3500	3600	3680	3700
3847	3876	3897	4208	4250	4500	4500	4612	4983	5000	5000	5000	5000	5100
5306	5800	8641	9875	12621	30254	46656							

## **Appendix F: Code PHP/HTML of experiment page**

The PHP/Html code as well as the data files containing the estimates of the four groups can be accessed in the following address:

<http://www.di.ciencias.ulisboa.pt/~lcorreia/WoC/>

## Appendix G: Experiment layout of Group 0

SURVEY

Thanks for taking part on this survey, it only takes 2 minutes!

Win a 10\$ Bonus if your answer is one of the 3 best!



QUICK RULES:

- 1) Please try your best to be honest
- 2) You can only take this survey once!
- 3) After you have submitted your answer and your MTurk ID you will get a code to paste on the MTurk site.

Look at the images below, how many jelly beans exist in the jar?

Type in your guess

Only numbers



Best 3 answers win a 10\$ bonus

We will compare your guess with the actual amount of jelly beans in the jar and if your answer is one of the 3 closest we will transfer a 10\$ reward to your MTurk account.

So make sure you give your best shot!

Finish

Write Your MTurk ID

MTurk ID

Submit

ISCTE/ Faculdade de Ciencias da Universidade de Lisboa

## Appendix H: Experiment layout of Group 1

SURVEY

Thanks for taking part on this survey, it only takes 2 minutes!



Win a 10\$ Bonus if your answer is one of the 3 best!

QUICK RULES:  
1) Please try your best to be honest  
2) You can only take this survey once!  
3) When you submit your answer you'll get a code to paste at MTurk site.

Look at the images below, how many jelly beans exist in the jar?

Type in your guess

Only numbers



Hint

If you want, you can take into consideration the guesses of other participants.  
Based on all the guesses of other participants, the closest guesses so far are  
(in no particular order) 3154. . . 3136. . . 3138. . . 3129. . . 3123

Best 3 answers win a 10\$ bonus

We will compare your guess with the actual amount of jelly beans in the jar and if your answer is one of the 3 closest we will transfer a 10\$ reward to your MTurk account.  
So make sure you give your best shot!

Finish

Write Your MTurk ID

MTurk ID

Submit

ISCTE/ Faculdade de Ciencias da Universidade de Lisboa

## Appendix I: Experiment layout of group 2

SURVEY

Thanks for taking part on this survey, it only takes 2 minutes!

Win a 10\$ Bonus if your answer is one of the 3 best!


QUICK RULES:

- 1) Please try your best to be honest
- 2) You can only take this survey once!
- 3) After you have submitted your answer and your MTurk ID you will get a code to paste on the MTurk site.

Look at the images below, how many jelly beans exist in the jar?

Type in your guess

Only numbers



[Hint](#)

If you want, you can take into consideration the guesses of other participants.

Based on all the guesses of other participants, the guesses between 200 and 4380 were the most common.

Best 3 answers win a 10\$ bonus

We will compare your guess with the actual amount of jelly beans in the jar and if your answer is one of the 3 closest we will transfer a 10\$ reward to your MTurk account.

So make sure you give your best shot!

Finish

Write Your MTurk ID

MTurk ID

Submit

ISCTE/ Faculdade de Ciencias da Universidade de Lisboa

## Appendix J: Experiment layout of Group 3

SURVEY

Thanks for taking part on this survey, it only takes 2 minutes!

Win a 10\$ Bonus if your answer is one of the 3 best!

QUICK RULES:

- 1) Please try your best to be honest
- 2) You can only take this survey once!
- 3) When you submit your answer you'll get a code to paste at MTurk site.



Look at the images below, how many jelly beans exist in the jar?

Type in your guess

Only numbers

[Hint](#)

If you want, you can take into consideration the guesses of other participants.



Best 3 answers win a 10\$ bonus

We will compare your guess with the actual amount of jelly beans in the jar and if your answer is one of the 3 closest we will transfer a 10\$ reward to your MTurk account.

So make sure you give your best shot!

Finish

Write Your MTurk ID

MTurk ID

Submit

ISCTE/ Faculdade de Ciencias da Universidade de Lisboa