The More the Merrier? Understanding the Effect of Group Size on Collective Intelligence

By

Nada Hashmi

M.S. Computer Science Studies
University of Maryland, 2005

S.M Engineering and Management Science
Massachusetts Institute of Technology, 2008

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Signature redacted

Signature of Author: _____
Department of Management
August 11, 2017

Signature redacted

Certified by: _____
Thomas W. Malone
Patrick J. McGovern Professor of Management
Thesis Supervisor

Signature redacted

Accepted by: _____
Catherine Tucker
Sloan Distinguished Professor of Management
Professor of Marketing
Chair, MIT Sloan PhD Program

The More the Merrier? Understanding the Effect of Group Size on Collective Intelligence

By
Nada Hashmi

Submitted to the Sloan School of Management on August 11, 2017 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Management

Abstract:

This dissertation explores how group size affects collective intelligence. It is composed of three quantitative studies. The first study explores how time pressure in small groups (size 4) and large groups (size 20) affected collective intelligence. The results showed that the large groups significantly and consistently outperformed the small groups in different time pressure conditions. This led to the second study which explored whether the collaboration tool used in the first study might have provided unexpected benefits for large groups that counteracted any process loss in the large groups. While the results from the second study confirmed that the collaboration tool did indeed significantly improve the collective intelligence score of groups, one surprising result was that this effect occurred, not only in large groups (size 20), but also in small ones (size 4). The final study then set out to explore this surprising result in further detail by including a variety of group sizes (sizes 5, 10, 15, 20, 25, 30, 35 and 40) in both the collaboration conditions. It was hypothesized that by including more group sizes, the study would determine whether a curvilinear (inverted-U) relationship existed. The results not only confirmed the curvilinear (inverted-U) relationship but also suggested an optimal group size of about 30 for groups with the collaboration tool and 25 for groups without the collaboration tool.

Thesis Committee:
Thomas W. Malone (Chair)
Patrick J. McGovern Professor of Management

Lotte Bailyn
T Wilson (1953) Professor of Management, emerita
Anita Williams Woolley

Assoc Professor of Organizational Behavior & Theory at Tepper School of Business in Carnegie Mellon University

Acknowledgements

In the Name of God, The Most Gracious, The Most Merciful

I dedicate this thesis to my parents: Dr. Nasim A. Hashmi and Aisha Hashmi, my pyaray Abbu Ammi.

When I think of my journey, the people who helped me, the circumstances that brought me to this position, I wonder in amazement at how beautifully and coincidently everything came together. Yet I realize that nothing happens by chance or coincidence and so I must start by acknowledging and thanking God first and foremost – for blessing me with the right people, circumstances and ability to complete the PhD at MIT.

I thank my advisor, Thomas W. Malone, for his unwavering guidance, support and patience throughout my PhD. Thank you so much for believing in me, encouraging me and challenging me at every step of the way. Thank you so much for being an exceptionally brilliant and talented Professor and a genuinely authentic caring individual. It has been an absolute honor and enriching learning experience being your student. I thank my committee members, Lotte Bailyn and Anita Williams Woolley. Thank you so much for being there, your invaluable expertise and the constant pep talks. I thank the many collaborators of the Center of Collective Intelligence: David Engel, Young Ji Kim, and the countless UROPS who helped throughout. I thank all the software engineers from POD consulting for their hard work in developing the software needed for my thesis. I thank the administrative staff at MIT, specifically Richard Hill and Liz McFall, for going out of their way to set up some of the meetings and working with me side by side.

Last but not the least, I thank my family and buzoorgs: Abbu, Ammi, Sahar, Saira, Ibrahim, Maliha, Basma, Bibi Jan, Mian Bhai, Qadri Uncle and Kamal Uncle. Thank you for being my inspirations, for your unending love and for all your support.

Finally, I would like to send my sincerest and deepest salutations, peace and God's Blessings upon Sayidna Muhammad Peace Be Upon Him, his family and companions, Peace Be Upon Them All. Sallo Alan Nabi Sallalaho Alayhi Wa Salam.

Contents:

Chapter 1: Background

Collective Intelligence

Historically, there are millions of examples of groups coming together to accomplish seemingly impossible tasks and challenges. Whether it's how ants come together to create phenomenal structures underground or how hundreds of individuals come together to create operating systems, one thing is clear – 'collective intelligence' is not new and has existed for a very long time. What is relatively new is how researchers are defining what is collective intelligence (Hiltz and Turoff 1978, Smith 1994, Malone and Bernstein 2015), creating metrics to measure it (Wooley et el 2010) and understanding how it manifests itself in different fields such as in biology (Gordon 2015), cognitive diversity (Aggrawal et el 2015) and gender diversity (Woolley and Malone 2011). My research, in particular, explores the effects of group size on collective intelligence and aims to find optimal group sizes.

In terms of defining collective intelligence, while many researchers have defined collective intelligence differently, my thesis uses Malone and Bernstein's (2015) simple definition: *"groups of individuals acting collectively in ways that seem intelligent'*. Specifically, in their definition, they chose not to define 'intelligence', brought focus on the 'act' of working together and not the product produced by the act and finally, said that it is collective intelligence if the observer deems it to be. This definition is inclusive and allows the observer to decide if the phenomenon is an act of collective intelligence. Finally, this definition does

not judge an act of collective intelligence by its performance. That is, even if the hundreds of individuals came together and created a 'bad' product, it could still be deemed as an act of collective intelligence.

Other definitions include:

- 'A collective decision capability [that is] at least as good as or better than any single member of the group'. (Hiltz and Turoff 1978)
- 'A group of human beings [carrying] out a task as if the group, itself, were a coherent, intelligent organism working with one mind, rather than a collective of independent agents' (Smith 1994)
- 'A form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills' (Levy 1994)

Hiltz and Turoff (1978) limit their definition relative to the performance of a single individual. However, as stated, there may be cases where the single individual performs better than the group of individuals. An example would be where one of the individuals in the group is an expert in math but the group refuses to recognize their expertise. Hence, as a group, they may perform worse than the expert in the group. However, it is still a 'collective decision' but perhaps a bad 'collective decision'.

Smith (1994) limits the definition by requiring the task be completed by 'human beings'. Hence, any task completed by animals or humans and machines together would not be regarded as collective intelligence. Levy (1994) requirement for real time is a limiting factor that excludes asynchronous group work. Furthermore, that definition raises more questions than it answers, such as what is universally distributed intelligence and re-opens the debate on another complex concept of intelligence.

Intelligence, on its own, has been a concept that is difficult to objectively define and one that is confounded by many debates and controversies. Despite its complexity, many researchers have attempted to capture this concept. Wechsler, a student of Spearman who developed the most widely used method for measuring intelligence, defined intelligence as "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment" (Wechsler, 1944). Collectively, the American Psychological Association used the idea that "individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, [and] to overcome obstacles by taking thought" to formalize their definition of intelligence (Neisser et al., 1996).

Legg and Hutter (2006) compared and contrasted over 50 different definitions from prominent researchers in the field of intelligence and found definitions to be differing in terms of being concise, precise or more general. The common features, however, amongst all those definitions were that intelligence:

- "[Is] a property that an individual agent has as it interacts with its environment or environments."
- "Is related to the agent's ability to succeed or profit with respect to some goal or objective."
- "Depends on how able the agent is to adapt to different objectives and environments."

Thus, one can summarize intelligence as the general mental ability of an individual to learn and apply knowledge to manipulate the environment, as well as the ability to reason and have abstract thought.

Furthermore, like the definitions of intelligence, the measurement and operationalization of intelligence has also been surrounded by controversy and disagreement. While there are a number of different methods for measuring intelligence, the standard and most widely accepted is by measuring an individual's general cognitive ability (*g*). This method was first developed by Spearman and is considered the 'g factor' (Spearman, 1904).

The 'g factor' explains the variance across many different cognitive abilities. Deary (2000) described it when stating that "human cognitive abilities form a multi-level hierarchy, with a general factor (Spearman's g) at the pinnacle, many very specific abilities at the bottom and separable but correlated group factors of ability in between at one or more levels". More

specifically, it is the first factor that emerges from factor analysis conducted on tasks. These tasks represent different cognitive abilities and have positive inter-item correlation. This first factor generally accounts for more than 30% of the variance and approximately twice as much as the second factor.

From this, researchers have gone on to develop many different tests to measure intelligence, the 'intelligence quotient' or IQ. A typical intelligence test includes items from various cognitive domains and draws on basic capacities and abilities, such as short-term memory, verbal knowledge, spatial visualization and perceptual speed. To standardize an IQ test, a representative sample of the population is tested using each test question. The super-set of questions is narrowed down to a smaller subset of questions that elicit a normal distribution of responses and that reliably predict performance on a broader range of tasks or cognitive abilities. Individual Intelligence has been researched for decades and despite the controversies surrounding it, the remarkable finding by Spearman in 1904 of measuring it and finding that an individual who does well on one mental task tends to do well on most others continues to hold today. (Spearman, 1904)

This research applies the concept of measuring individual intelligence to measuring collective intelligence. Specifically, in previous work by Woolley, Chabris, Pentland, Hashmi and Malone (2010), the group defined the metric of collective intelligence of a group as the "general ability of the group to perform a wide variety of tasks". In that study, my colleagues

and I devised a way to measure 'collective intelligence'; the premises for measuring collective intelligence came from the theoretical framework of individual intelligence.

Furthermore, we showed that the predictive power from collective intelligence was above and beyond what can be explained by knowing the abilities of the individual group members. We showed there is something inherent about the group itself, that is, the group dynamics, that is responsible for the performance.

The collective intelligence metric was developed using the same approach used in measuring individual intelligence. From that, we found evidence of collective intelligence that explained a team's performance on a wide variety of tasks. After conducting a factor analysis, the first factor that emerged explained 43% of the variance from the first study and 44% from the second study. In both the studies, the second factor explained less than half of the variance explained by the first factor. Finally, this first factor was also positively correlated with the diverse set of tasks, including the criterion task. This indicated that the collective intelligence, 'c', predicted the future performance of the team as well. Furthermore, this was not very strongly correlated with the average general mental ability (general intelligence) of the individual team members or of the team's most intelligent member. It was not correlated with the group satisfaction, group psychological safety, team cohesion, internal motivation or individual personality. It was, however, correlated with the social sensitivity of the group, equality in distribution of conversational turn-taking and the proportion of female in a group (mediated by social sensitivity).

This finding has led to a whole new range of questions. Amongst the questions that intrigued me, understanding what effect group size has on collective intelligence stood out. Hence, I conducted three studies:

- Effect of Time Pressure and Group Size
- Effects of Extreme (Small versus Large) Groups Sizes
- Effects of a Range of Group Sizes

All three of my studies used the Platform for Online Group Studies (POGS) and five tasks (brainstorming the uses of a brick, solving a Sudoku puzzle, solving abstract reasoning questions, reproducing text and unscrambling words) with subjects from Amazon Mechanical Turk. Hence, my thesis is organized as follows: First I start with a literature review of group size as related to my thesis. Then I describe the POGS system, the five tasks, and Amazon Mechanical Turk. The next three chapters are my three different studies. I conclude in the final chapter with a discussion and future implications.

Group Size

Jeff Bezos, CEO and Founder of Amazon.com, has a simple rule: if a team cannot be fed by two pizzas, then that team is too large. On the other hand, Cloudera has 70+ Project

Managers and 100+ developers all working together on one project Hadoop. Arguably, both are success stories.

As such, despite over 50 years of work on group sizes, the jury continues to be out on what is an 'optimal' group size. The number 5, however, has repeatedly appeared in different research scenarios. Hackman and Vidmar (1970) experimented with group sizes 2 through 7 to assess the impact of group process and performance over a variety of tasks. After completing the tasks as a group, individuals were asked to rate their member satisfaction and whether they thought the group was too large or too small. Few groups in dyads thought the group was too large while few groups thought groups of 7 were too small. From the intersection of those group sizes, Hackman and Vidmar arrived at the optimal group size of 4.6 (or 5 individuals). However, by the time Hackman published his book (2002), his rule of thumb for optimal group size was 6.

Mueller (2012) argued the optimal group size largely depended on the task at hand. To clean a stadium, 30 janitors were better than 5. However, in tasks that required greater coordination, she agreed with the number 6. She studied 212 workers within 26 teams in 7 different companies spread across three different industries. In this study, the groups were completing complex tasks in groups that ranged from group sizes of 3 to 19. Individuals were pinged by surveys daily during the work week as well as at the project midpoint. She measured the performance using the companies' metrics and found larger teams performed worse than smaller ones. Her main conclusion supported Ringelmann's theory (1913) that as

group size increases, individuals tend to social loaf; that is, put in a lot less effort. She reasoned that this was a result of relational loss; that is the tendency that an employee perceives that support, help and assistance is less available within the team as team size increases (Lakey and Cohen 2000). In later interviews discussing the work, she reiterated the optimal team size to be 6; after which relational losses negatively impact the performance.

Blenko et al. (2010), on the other hand, have argued the optimal group size to be 7. They argue this number is great for office groups that use spreadsheets and powerpoints and need to make decisions together (versus manual labor tasks). Each additional person after 7 people reduces decision effectiveness by 10% and groups with 17 or more members rarely make decisions other than when to take a lunch break.

Mao et al (2016) found the largest group size they studied to have the best performance. They studied individuals and groups of size 2, 4, 8, 16 and 32 and their performance in a complex task. In particular, individuals formed groups and engaged in 'crisis mapping'; online collaboration to monitor, classify and map real-time information about affected populations in the midst of a humanitarian crisis. Information is aggregated and processed from a variety of social media including tweets, Facebook, Instagram, etc. Mao et al simulated the December 2012 Typhoon Pablo that struck in the Philippines. They compared the results of the subjects they recruited benchmarked against the external results that occurred at the time of the disaster. For their experiment, they recruited from Amazon Mechanical Turk.

They had 18 groups of size 1, 11 groups of size 2, 6 groups of size 4, 4 groups of size 8, 4 groups of size 16 and 4 groups of size 32. Even though they found marginal performance to be decreasing as group size increased, the largest teams still outperformed all other teams overall. While they did not conclude that 32 is the optimal group size, their research countered findings that smaller groups of 5, 6 or 7 would have greater performance for complex tasks that required coordination and communication.

It is important to note, Mao et al's study was conducted virtually, that is using the internet as a tool, whereas the previous studies were conducted face to face. Furthermore, it employed resources and subjects familiar with completing work on the internet. This use of online collaboration tools may remove many of the physical impediments of working together that exist in face to face groups. Amongst their conclusions, they also emphasized that individuals benefit from collaboration that occurs from coordination virtually. Implicitly, this may be due to collaboration tools inherent in the use of an online tool. This is further discussed later in the chapter.

Aside from the Mao study, in general, the research mentioned thus far has argued that larger teams perform worse. Other research supports this claim as larger teams spend less effort (Latane, Williams & Harkins 1979), assume less responsibility (Wicker and Mehler 1971) and generally perform worse than individuals on smaller teams (Liden et al, 2004). Furthermore, as group size increases, member satisfaction decreases (Stiener 1972). In general, this is said to be a result of communication and coordination decreasing (Hare 1952, Hawkins 1962,

Kelley and Thibault 1954, Schneider and Zimet 1969) or members feeling threatened and being inhibited (Gibb 1951, Berkowitz 1958). Furthermore, process losses, that is, the costs of coordination and communication, increased in larger groups (Steiner 1972).

In particular, Steiner (1972) defined process loss as *a reduction in performance effectiveness or efficiency caused by actions, operations or dynamics that prevent the group from reaching its full potential*, including reduced effort, faulty group processes, coordination problems, and ineffective leadership. He measured it as 'actual productivity' being equal to 'potential productivity' minus 'process loss'. Steiner found large groups suffered from process losses greater than the smaller groups. Other researchers defined other process losses. Diehl and Stroebe (1987) defined production blocking that occurs as people take turns to relate their ideas. In larger groups, this turn-taking process prevents many ideas from being discussed or even being mentioned. They also discussed evaluation apprehension which is when individuals withhold ideas as they fear a negative reaction from other participants. Again, a process loss that effects larger groups more than small groups. All these reasons have supported the research stream in which large groups perform worse.

However, there does exist a research stream that supports large teams outperforming small teams. These studies noted as group size increases, productivity and performance increases as well (Fink and Thomas 1963, Fox et al 1953). Furthermore, larger groups bring more diverse solutions (Wanous and Youtz 1986). This is largely due to larger groups having access to more resources (Hare 1952), retaining more information (Horowitz and Bordens 2002).

To resolve the conflicting results where certain studies supported large teams while other studies supported small teams, many researchers proposed the effect of group size on performance is mediated by the type of task at hand (Anderson and Frank 1971, Littlepage and Silbiger 1992, Shaw 1976, Steiner 1972). This research used Steiner's taxonomy of group tasks (1972) that includes five different types of tasks: Additive, Compensatory, Disjunctive, Conjunctive and Discretionary. An additive task is one in which each individual's input is added together (e.g. shoveling snow, cleaning a stadium). A compensatory task is where individual inputs are averaged together (e.g. guessing the weight of a person). Disjunctive tasks are where the group selects one or more of the individual solutions by group members (e.g. solving a math problem). Conjunctive tasks require all the individual to contribute for the task to be complete (e.g. climbing a mountain while group members are tied to each other). A discretionary task is when a group decides how individual inputs relate to the task (e.g. vote for the best answer).

Gallupe et al (1992) found that larger groups outperformed smaller groups for additive and compensatory tasks but only if social loafing and production blocking could be prevented using collaboration tools. Yetton and Bottger (1983) confirmed these results but added only if the group used correct decision making to avoid bias. Hare (1976) and Stewart (2006) noted for conjuctive tasks, there was pronounced coordination and motivational loss in larger teams that decreased their performance. From these studies and others (Hackman and Vidmar 1970, Mueller 2012, Slater 1958), it is safe to conclude larger teams will

outperform smaller teams if social loafing is prevented, if they coordinate their work effectively and if they feel connected to each other. In small groups, this occurs naturally while the need to facilitate this through collaboration and communication tools becomes increasingly important for large groups. Hence, communication and coordination tools become vital in larger groups (Stasser and Titus 1985, Woolley et al. 2010, Gallupe et al., 1992, Valacich, Dennis and Connolly 1994).

Researchers have also tried to improve large group performance by using collaboration tools such as multi-channel chatrooms, shared online documents, social media, electronic voting systems and real-time feedback systems. Multi-channel chatrooms were found to assist creativity and result in more diverse solutions as they helped prevent production blocking (Thompson 2003). Shared online documents assist in preserving organizational memory and allow equal contribution as members are able to work on the document simultaneously. The use of social media within groups, similar to Facebook, Twitter, etc, allows for the expression of social cues that might otherwise be lost in large groups. Researchers found the use of social media within companies to help create awareness, assist in task/meeting coordination and idea generation/discussion (Riemer, Richter, and Bohringer 2010, Riemer and Richter 2010). Electronic voting systems help the best idea to percolate to the top and potentially the 'best answer' is selected (Hertel et al. 2005). Finally, real-time feedback systems that bring about awareness of individual participation levels by keeping track of contributions and level of communication help engage all the group members and prevent social loafing (Karau and Williams 1993).

Finally, another factor that affects group performance is time pressure. Literature studying time pressure on groups found increased time pressure decreases performance quality, forces members in a group to stop considering multiple alternatives, pressures members to engage in erratic and incomplete processing of information and refrain from constructive criticism (De Dreu 2003, De Grada et al. 1999, Durham et al., 2000, Karau and Kelly 1992, Kelly et al. 1997, Kelly and Karau 1999, Kelly and Loving 2004, Kruglanski and Freund 1983). That is, with time pressure, group performance decreases. Hence, my first study delved deeper into studying time pressure and group size.

Chapter 2: General Methodology

Platform for Online Group Studies

The studies by Woolley et al (2010) used a multitude of tasks that were pen-and-paper based. This required the subjects to be present face-to-face physically. In total, the complete battery of tasks took over five hours to complete. In addition, there was excessive administrative time required before and after each session to prep and score the tasks (Wooley et el, 2010).

Since then, MIT, Carnegie Mellon University and Union College collaborated to develop an online open source software system, Platform for Online Group Studies (POGS), which replicated similar tasks electronically and allowed subjects to log in remotely at the same time to complete the tasks together. POGS was also created with the flexibility to configure, develop and deploy new tasks in any order. Researchers are able to configure each session to specify the tasks, the order of the tasks as well as whether all the subjects or just one of them type in the answers in the workspace. Figure 1 below shows what a subject typically sees.

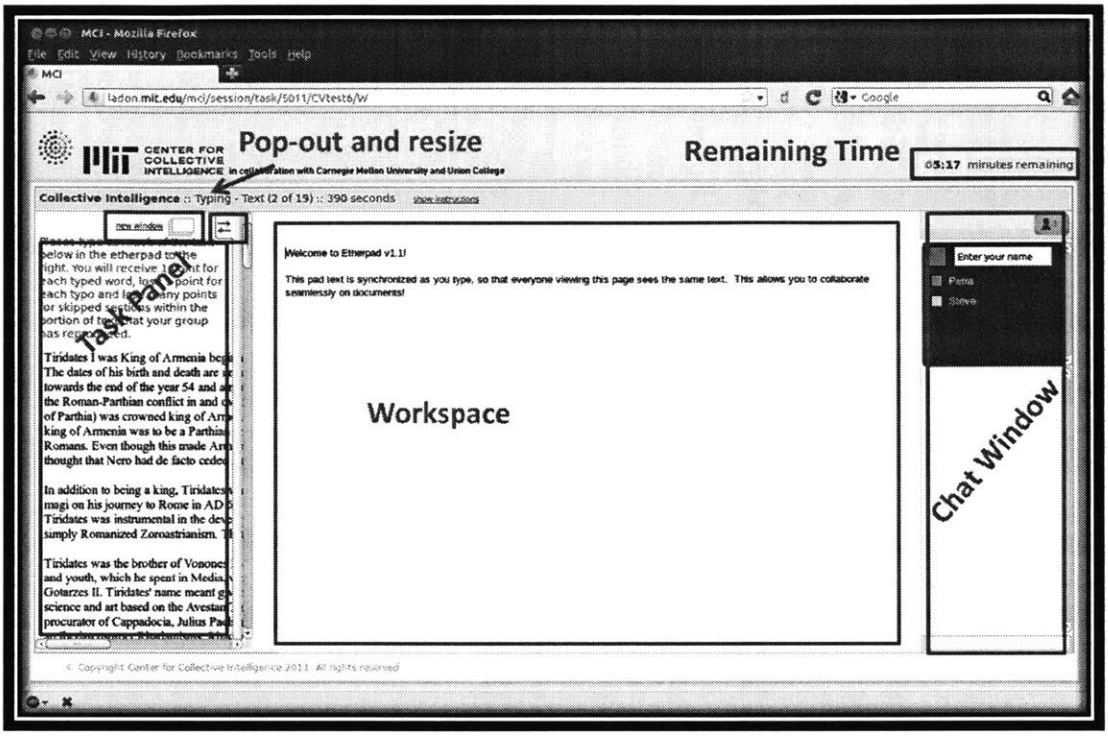


Figure 1: Typical layout of POGS from the perspective of a Subject

In each session, subjects log in with preassigned usernames and then are guided through the tasks the researcher has configured. Last tested, POGS was capable of handling four to five simultaneous sessions; one of which can contain up to 40 subjects. The largest successful configuration has been four simultaneous groups completing the experiment with the following subjects: 40, 15, 10 and 5 subjects.

In a number of studies, POGS has shown that the electronic version of the battery of tasks produced similar and equivalent results as the face-to-face studies (Engel et el 2014, Engel et el 2015, Aggarwal and Woolley 2013). POGS was used for all three of the studies in this

thesis. There were notable differences in how POGS was configured in each of the studies. Each of those differences are described in detail in the method section of each study.

Tasks used in POGS for this Thesis

For this thesis, five main tasks were used in the battery of tasks: reproducing text, solving abstract reasoning questions, unscrambling words, solving a Sudoku puzzle and brainstorming the uses of a brick – in that order. These tasks tested for competency across a variety of cognitive abilities that fit into each of the four major quadrants within McGrath's Task Circumplex (McGrath, 1984).

Reproducing Text

For this task, the subjects were shown text and asked to reproduce it in the workspace. They were given 6.5 minutes to complete this task. They received one point for each typed word and lost 1 point for each typo. This included repeating text and/or wrongly ordered text. Finally, they did not lose points for not typing the complete text but missed sections in the text did result in lost points.

Solving Abstract Reasoning Questions

One of the collaborators from Union college developed a series of abstract reasoning questions in which subjects were shown patterns and asked to select the next one from a

set of answers. The level of difficulty increases as the questions progress. This was similar to the Ravens Matrices test (Raven 1936). Altogether there are 18 questions, and subjects are allowed 3.5 minutes to answer them. Figure 2 below shows an example question.

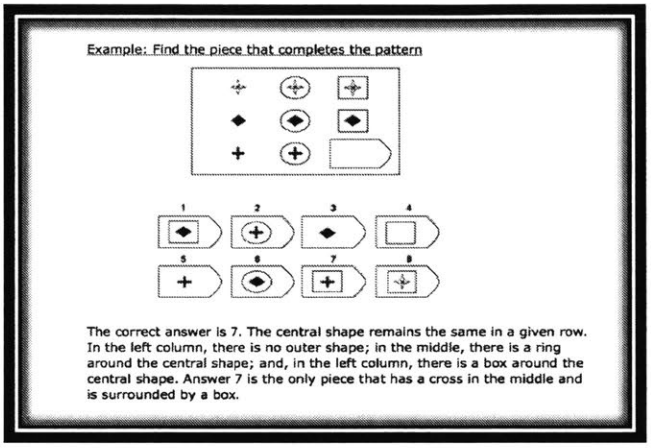


Figure 2: Example from Abstract Reasoning Questions

Unscrambling Words

A total of 24 scrambled words are given to the subjects to solve. For example, 'rpepa' unscrambled is 'paper'. The group has a total of 2 minutes to solve the unscrambled words.

Solving a Sudoku Puzzle

Sudoko is a logic-based, combinatorial number-placement puzzle where there is a 3x3 grid and in each cell, another 3x3 grid. Each sub 3x3 grid needs to have the numbers 1 – 9 only once while each row and column in the main 3x3 grid can contain only one instance of each

of the 9 numbers. Subjects are given a partially completed puzzle and then asked to complete the complete puzzle. They have a total of 3.5 minutes to complete the task. Figure 3 shows an example of the Sudoku puzzle.

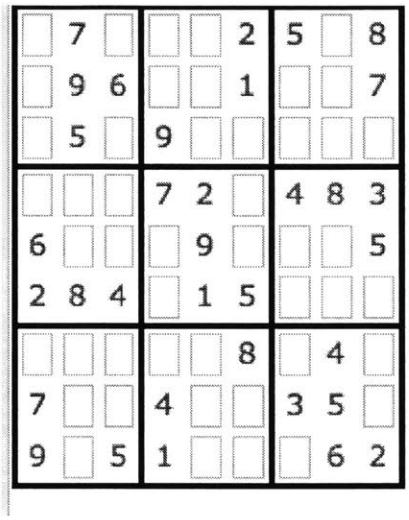


Figure 3: Example of a Sudoku Task used in POGS

Brainstorming uses of a 'Brick'

In this final task within the battery of tasks, the subjects are asked to brainstorm as many different uses for a brick as possible. Each line should be for one idea. Points are deducted for repeated ideas while additional points are awarded for creative and unusual uses. Subjects are given 2 minutes to complete this final task.

Session Configuration

POGS allows for each session to be configured individually. For the three studies, two different configurations were used. In one type of configuration, POGS managed how many groups for a particular group size it could create for each session. That is, if 20 people showed up and I specified groups of 5 subjects, POGS would automatically create 4 groups of 5 subjects each. However, if 24 subjects were to log in, POGS would still create 4 groups of 5 subjects each. The remaining 4 subjects would be still be provided a code to receive payment and told that we do not require them to complete the experiment. This configuration is useful and practical when there were few group comparisons required. However, this set up did not allow different group sizes to run together in a session and it created wasted resources.

In the second type of configuration, I specified for each session the number of groups and how many in each group depending on the number of subjects that would log in. That is, if 20 subjects were to log in, I could specify to POGS to create groups with the following number of subjects in each one: 10, 5, 3, 2. In total, four groups with varying size would be created and run in this session. In the same way, if an odd number of subjects logged in, for example 21, I could specify five groups: 10, 5, 3, 2, 1. In this way, there was a greater degree of flexibility in how many groups of varying sizes and fewer resources are wasted.

Given the limitations of the first configuration, the second configuration was implemented after the first study. Hence, the first study used the first configuration while the remaining studies used the second configuration.

Amazon Mechanical Turk

Background

Until recently, researchers relied heavily on student subject pools or professional Internet panels from companies, such as Qualtrics, to collect data. While student subject pools faced the threat of external validity and were a younger population who were geographically constrained (Peterson and Merunka 2014), professional panels suffered from high costs, slower data collection as well as significant time and effort administratively (Brandon et al., 2014). Amazon Mechanical Turk (MTurk) launched in 2002 and allowed flexible, affordable and immediate access to subjects. While its original intent was purely crowdsourcing for menial tasks, it grew rapidly and expanded – quickly becoming a one-stop shop for getting all types of remote online work done, creating tasks, and survey data collection (Potin 2007). This marketplace provided *workers* (Turkers), with thousands of 'Human Intelligence Tasks (HITs)' that they can choose to complete for pay at their convenience and *requesters* with the ability to offer unlimited HITs. In 2014, it was noted that MTurk had over half a million workers from all over the world (Marvit 2014). On an average day, there are more than 430,000 HITs available on MTurk of which academic research studies account for less than 10% at any given time (Sheehan and Pittman 2016).

MTurk provided the marketplace for both workers and requesters. It also provided mechanisms to track *qualifiers* of the Turkers. These qualifiers range from number of total HITs completed, percent of successful completions and location. This allows the requesters to select serious Turkers. MTurk, however, did not provide similar ratings or qualifications for requestors and provided very little policing. As a result, complementary websites quickly sprung up to assist Turkers develop social networks and allow Turkers to report, rank and/or avoid bad HITs or requesters (such as requesters who refused to pay even after the work was completed). In particular, TurkOpticon by University of California, San Diego, TurkNation, TurkGrind and many others have ranking systems for requesters, online discussion forums and other tools for Turkers. This and other forums aimed to helping requesters have created a whole ecosystem around MTurk.

MTurk Characteristics

In general, Turkers tend to be female, around 30 years old, 'overeducated, underemployed, less religious and more liberal' than the general US population (Berinsky, Huber, & Lenz, 2012; Paolacci et al., 2010; Shapiro, Chandler, & Mueller, 2013). Self-reporting methods show that Turkers have both extrinsic (e.g. 'making ends meet') and intrinsic motivations (e.g. 'tasks are fun') (Ross et al., 2010). Personality studies found that Turkers are less extraverted subjects and more socially anxious than the student subject pools (Shapiro et al., 2013) and there is some evidence that Turkers might even be less emotionally stable than the general population (Goodman et al., 2013; Kosara & Ziemkiewicz, 2010).

In 2010, Ipeorotis found 46.8% of the Turkers were from the US and 34% from India (Ipeorotis 2010). Furthermore, from his survey, he found most workers spent between 4 – 8 hours and completed on average 20 – 50 HITs on MTurk each week. The average MTurker he surveyed showed an earning between $1 - $5 per week. The survey motivated him to create a real-time tracker, www.mturk-tracker.com, where, at any given time, anyone can view the number of HITs available, Turker characteristics as well as top requesters (Ipeorotis 2010; Difallah, Catasta et al. 2015).

Finally, the average pay for a HIT on MTurk depends on the complexity of the task and many blogs have pointed to using the federal minimum wage as the expected pay ($7.25/hour). However, in reality, it could be as low as $2.25/hour (Berinsky, Huber, & Lenz, 2012).

Use of MTurk as a Subject Pool in Social Science Experiments

Several studies that compared student subject pools, professional Internet panels and MTurk found that MTurkers were more representative of the US population, had greater diversity (36% were non-white) and were older (Burhmester, Kwang et al. 2011; Kees et al. 2017; Smith et al. 2016; Mason and Suri 2012). Furthermore, Kees also specifically tested for data quality that included measurements of involvement (Kees, Burton, and Tangari 2010), attention checks (Smith et al. 2016), instructional manipulation checks (Oppenheimer, Meyvis, and Davidenko 2009) and measures of research participation (such as, time to complete the survey) and general computer knowledge. From amongst two different

student subject pools, a Qualtrics Professional Internet Panel, a Lightspeed Professional Internet Panel and a Turkers subject pool, he found that Turkers had the best reliability, fared best in manipulation checks, instrumental manipulation checks and attention checks. Kees also found that the quality of the data from MTruk was not correlated with the rate of pay. In addition, MTurk was the cheapest and fastest option amongst his comparison pool.

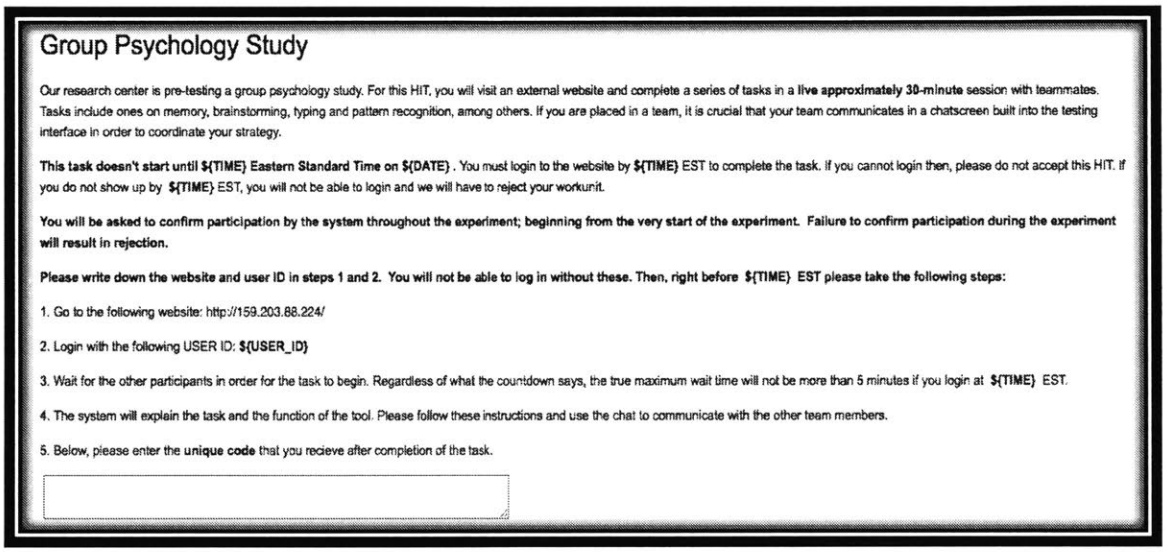
Despite Paolacci and Chandler (2014) having similar findings as Kees (2017), Paolacci still cautioned against the use of Turkers as subjects and insisted researchers must show complete transparency in the selection and type of Turkers used as subjects.

MTurk and Turkers for this research

MTurk was used to recruit subjects for all three of the studies in this thesis. While each study differed slightly for recruiting methods and criterion, there was a general overarching recruiting method. During pilot studies, I noted the difference between the number of Turkers who accepted the HITs and the ones who actually logged in and completed the studies. Hence, to develop the recruiting method and criterion that attracted reliable Turkers, I conducted a survey, engaged Turkers on TurkGrind as well as tested out and experimented with different criterion during pilot studies.

Recruiting Method

MTurk allows for 'batches' to be created and advertised at once. A 'batch' consists of a number of HITs that follow a template, but I can specify unique variables for each HIT inside of the template. For my studies, the general template for a batch is shown in Figure 4. This template was used across the three studies. The variables I specified for each HIT were the time, date and the user ID.



Group Psychology Study

Our research center is pre-testing a group psychology study. For this HIT, you will visit an external website and complete a series of tasks in a **live approximately 30-minute** session with teammates. Tasks include ones on memory, brainstorming, typing and pattern recognition, among others. If you are placed in a team, it is crucial that your team communicates in a chatscreen built into the testing interface in order to coordinate your strategy.

This task doesn't start until ${TIME} Eastern Standard Time on ${DATE}** . You must login to the website by **${TIME} EST to complete the task. If you cannot login then, please do not accept this HIT. If you do not show up by **${TIME}** EST, you will not be able to login and we will have to reject your workunit.

You will be asked to confirm participation by the system throughout the experiment; beginning from the very start of the experiment. Failure to confirm participation during the experiment will result in rejection.

Please write down the website and user ID in steps 1 and 2. You will not be able to log in without these. Then, right before ${TIME} EST please take the following steps:

1. Go to the following website: http://159.203.88.224/
2. Login with the following USER ID: **${USER_ID}**
3. Wait for the other participants in order for the task to begin. Regardless of what the countdown says, the true maximum wait time will not be more than 5 minutes if you login at **${TIME}** EST.
4. The system will explain the task and the function of the tool. Please follow these instructions and use the chat to communicate with the other team members.
5. Below, please enter the **unique code** that you recieve after completion of the task.

Figure 4: Recruiting Text Advertised on MTurk

As such, each MTurker was able to accept one HIT with the time, date and the user ID that was unique to them. While MTurk allowed an unlimited number of HITs to be published per batch, the upper limits of HITs I published in any one batch was 80. In general, I had a response rate between 70 – 80% from each batch for each study.

For studies 1 and 2, I would advertise up to 24 hours in advance for the study. However, the pilot studies showed a low rate of Turkers who signed up and who actually logged in. The rate ranged from as low as 40% to 50%. The comments from the survey and engagement with the Turkers revealed that most Turkers use MTurk ad hoc and at random. Many of them log into MTurk when they have lunch breaks or a few hours to spare. Hence, anything that requires a time too far into the future would decrease the probability that Turkers would actually login at the required time and 24 hours was too far in the future. I experimented between 5 to 2 hours for the pilot studies. Surprisingly a 2-hour window from the time the batch was advertised to when the experiment was run brought the highest response rates between 70 – 80%. This advertising window time was then used for study 3.

Recruiting Criterion

MTurk further allowed requesters to specify different demographics and characteristics that Turkers can qualify against. For this thesis, I specified the following criterion across the three studies:

1. Turker must be present in the US or Canada.
2. Turker must have completed at least 500 HITs.
3. Turker must have a success rate of 80% or more.

The first criteria was meant to prevent non-English speakers and avoid different cultural biases. While confining the geography to the US and Canada doesn't necessarily guarantee

that, it increases the chances of the subjects being English speakers as well as having similar cultural biases.

The second and third criteria aimed at attracting reliable Turkers. In addition to the survey and engagement with Turkers, I also consulted with Professor Panagiotis G. Ipeirotis who conducts extensive research using MTurk. From those sources, the criterion of Turkers completing at least 500 HITs as well as a success rate of 80% or more was selected.

Preventing Repeat Turkers

Another concern for studies on MTurk where each experiment required unique subjects and no repeats was how to prevent repeated Turkers. MTurk does not allow any direct mechanism to prevent repeat Turkers. It does however store the list of all the Turkers who have participated in HITS from each requester. Furthermore, it allows requesters to assign 'qualifiers' to each Turker. The original intent for 'qualifiers' was to allow requesters to specify the skill, reputation or ability required. Requesters would ask Turkers to take a test or survey and the results would determine the qualification which is any numerical value. Based on that, requesters were able to have Turkers with certain skills, abilities, etc.

However, requesters have put together a hack where they assign each Turker who has already participated in their study a qualifier, such as 1. After that, in addition to any other criterion a requester might require, the requester requests another criteria that the Turker

'NOT' have the assigned qualifier. Hence, only the Turkers who have not taken the study before are eligible. Hence, this system prevents repeated users.

This is the system I developed for all my studies and prevented repeated Turkers for my studies.

Chapter 3:

Study 1 - Effects of Time Pressure and Group Size

Introduction to Effects of Time Pressure and Group Size

As stated above researchers have found increased time pressure decreases performance quality (De Dreu 2003, De Grada et al. 1999, Durham et al., 2000, Karau and Kelly 1992, Kelly et al. 1997, Kelly and Karau 1999, Kelly and Loving 2004, Kruglanski and Freund 1983). However, I was not able to find any research to date that studied time pressure over a range of group sizes or compared large and small group sizes with respect to group size.

Furthermore, during pilot studies for POGS to test its upper limit of group sizes, I noted the subjects in larger group sizes in our experiments complained about not having enough time for the tasks. For example, I received emails from participants with the following messages:

> *"That was an interesting but also frustrating study simply because it is nearly impossible to collaborate with 12 people in such short time. I understand that is the point and I had fun though, thanks!" Participant email message*

> *"I just completed your very interesting job. If I could suggest one thing for you in the future. It would be that there were way to many people*

trying to talk at one time. Maybe have smaller groups. It made it very hard for every person to try and participate in this. But it was very fun." Participant email message

It is worth noting that the experiments did not require any feedback; Turkers felt strongly about this and they sent direct feedback. These email messages and the chat messages suggested that large groups did not have enough time and prompted me to explore varying the time.

I believe large groups have increased time pressure when they are expected to perform the same complex task in the same time length as a small group. This is because process losses and coordination complexities hinder effective group performance - they are expected to communicate, process more information (input from more members) and then try to have a consensus from more people.

Smaller groups do not face the same kind of process losses and should be able to communicate and coordinate more effectively and efficiently. It is conceivably possible that smaller groups may also benefit from more time; which reduces time pressure. However, at a certain point, more time will not provide smaller groups any more additional benefits.

Hence, increasing the time could help reduce the process losses that disadvantages the large group sizes. Therefore, I wanted to explore the effect of time and group size on the collective

intelligence scores. In particular, I wanted to explore whether the collective intelligence changes between large and small groups under different time lengths for each task.

Hypotheses

Given our rationale, we developed one hypothesis:

> *H1: "When the time allowed is short enough, small groups will outperform large groups."*

As the emails from the participants indicated needing more time to coordinate with larger groups, there was a strong intuition that coordinating in larger groups became more difficult. There may be other process losses, in addition to coordination loss, that increase as group size increases which negatively impact the group performance. This includes social loafing, production blocking, cognitive interference and evaluation apprehension (Ringelmann 1913, Dennis and Williams 2005, Steiner 1972). In addition, communication levels decrease (Hare 1952, Hawkins 1962, Kelley and Thibault 1954, Schneider and Zimet 1969) and that too leads to larger groups performing worse (Liden et al, 2004)

As such, we hypothesize that time pressure has a greater negative impact on larger groups as compared to small groups.

Method

Data and Setting

A total of 40 groups were run; there were 20 groups of group size 4 and 20 groups of group size 20. For this study, four different time lengths were considered:

Tx1: approximately 30 minutes

Tx1.5: approximately 45 minutes

Tx2: approximately 1 hour

Tx3: approximately 1.5 hours

The times varied by a few minutes give or take depending on how long before Turkers completed logging into POGS.

At each time length, the time was increased proportionally per task. For example, at Tx1, the typing task was equivalent to 4 minutes. At Tx1.5 was 6 minutes, Tx2 was 8 minutes and Tx3 became 12 minutes. As such, the total time for each session also followed the same logic.

For the first study, the criteria selected to qualify a Turker were:

- located in the US or Canada
- have successfully completed 1000 or more HITs
- have an approval rate of 95%

- never have completed our experiment before

In addition, I paid above the minimum wage ($3.50 at Tx1). Similarly, I paid $5.25 for Tx1.5, $7.00 for Tx2 and $10.50 for Tx3. These criteria allowed us to attract serious workers. In order to reach the larger group sizes I over-recruited with turn out rates being at just under 70% after sign up.

Once a Turker logged into POGS, Turkers were randomly placed in groups of 4 or 20 in one of the following conditions: Tx1, Tx1.5, Tx2 and Tx3.

I conducted a power analysis for a two-way ANOVA for 2x4 groups in G*Power using an alpha of 0.05, a power of 0.80, degrees of freedom at 3 and a large effect which resulted in showing 5 groups per condition was sufficient. As such the following conditions were formulated:

	Tx1	Tx1.5	Tx2	Tx3
Group size 4	5 groups	5 groups	5 groups	5 groups
Group size 20	5 groups	5 groups	5 groups	5 groups

For this study, all the experiment sessions were run during business hours in the weekday. The batches were released on Mechanical Turk 24 hours in advance. Each session consisted only of groups of 4 or groups of 20. There were no sessions that had both groups of 4 and 20. Sessions were alternated to run one group size (either 4 or 20) each day.

As this was the first time POGS was running larger groups, many technical errors occurred. Many of the technical errors were critical and prevented collecting data until the error was fixed. This also resulted in needing to repeat many of the sessions to collect clean data. As such, data collection for this study lasted for a period of five months.

Dependent Variable

The dependent variable in this study was the collective intelligence score. I calculated the collective intelligence score by running a factor analysis using the results of the five tasks above and verifying if indeed it represented 'collective intelligence'; that is, the first eigenvalue factor explained roughly more than 30% of all the variance, more than half of the second factor and there was positive inter-item correlation amongst the items.

Independent Variables

For this study, our independent variables were the group sizes and the time lengths. As explained above, in this study, there were two group sizes: group size 4 and group size 20. There were four time lengths; Tx1, Tx1.5, Tx2 and Tx3.

Results

The descriptive statistics for each group size at each of the different time lengths is shown below. Again, there were 5 groups per condition (time length and group size). Table 1 below shows the descriptive statistics for each for the time lengths and group sizes.

Table 1: Descriptive Statistics

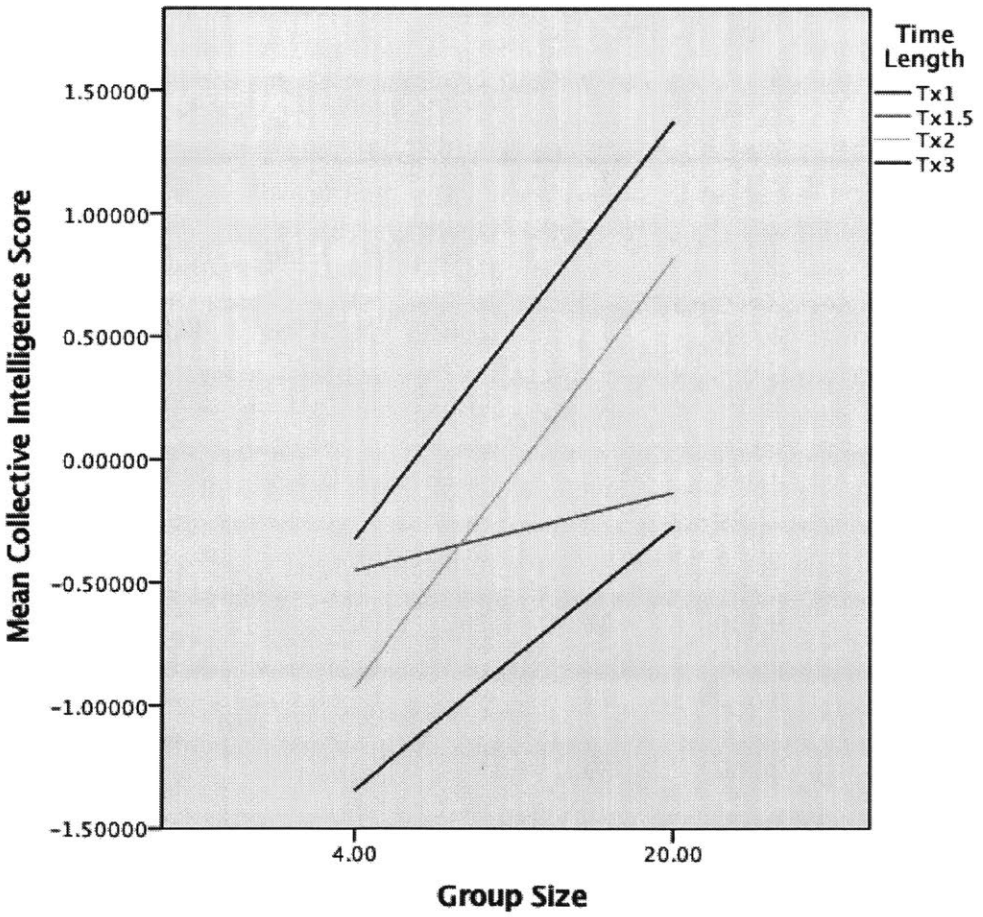
Dependent Variable: Collective Intelligence Score

Time Length	Group Size	Mean	Std. Deviation	N
Tx1	4.00	-1.3462522	.47089336	5
	20.00	-.2674283	.54569835	5
	Total	-.8068402	.74444322	10
Tx1.5	4.00	-.4502576	.61198755	5
	20.00	-.1310084	.62230776	5
	Total	-.2906330	.60571199	10
Tx2	4.00	-.9286946	.46143699	5
	20.00	.8215207	.53719915	5
	Total	-.0535870	1.03624119	10
Tx3	4.00	-.3212368	.66366581	5
	20.00	1.3756737	.33813450	5
	Total	.5272185	1.02295392	10
Total	4.00	-.7616103	.66100186	20
	20.00	.4496894	.84566237	20

Total	-.1559604	.96823615	40

A visual plot of the collective intelligence scores for each group at the four different time lengths showed that the large groups consistently had a higher collective intelligence score than the small groups. Graph 1 shows the results of mean scores of the collective intelligence at each time length and group size.

Graph 1: Results of the effects of Time Pressure and Group Size



A two-way ANOVA was conducted that examined the effect of time length and group size on the collective intelligence score. Group size and time length each differed significantly, F(1,32)=50.213, p=0.000 and F(3, 32)=10.484, p=0.000 respectively. The interaction between the effects of group size and time length was statistically significant, F(3, 32) = 3.820, p = 0.019. Table 2 below shows the results.

Table 2: Tests of Between-Subjects Effects

Dependent Variable: Collective Intelligence Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	27.211[a]	7	3.887	13.304	.000	.744
Intercept	.973	1	.973	3.330	.077	.094
Time_length	9.190	3	3.063	10.484	.000	.496
Group_size	14.672	1	14.672	50.213	.000	.611
Time_length * Group_size	3.349	3	1.116	3.820	.019	.264
Error	9.350	32	.292			
Total	37.535	40				
Corrected Total	36.562	39				

a. R Squared = .744 (Adjusted R Squared = .688)

A pairwise comparison of each time length shows that, aside from Tx1.5, large groups

significantly differed from the small groups. At Tx1.5, there was no significant difference between the CI scores of large and small groups. Table 3 below displays the complete results.

Table 3: Pairwise Comparisons of Mean Differences

Dependent Variable: Collective Intelligence Score

Time Length	Mean Difference (Group size 4 – Group size 20)	Std. Error	Sig.	95% Confidence Interval for Difference	
				Lower Bound	Upper Bound
Tx1	-1.079*	.342	.003	-1.775	-.382
Tx1.5	-.319	.342	.357	-1.016	.377
Tx2	-1.750*	.342	.000	-2.447	-1.054
Tx3	-1.697*	.342	.000	-2.393	-1.001

Discussion

H1: "When the time allowed is short enough, small groups will outperform large groups"

Unsupported. The results showed that, at the chosen time lengths, the large groups consistently outperformed the small groups significantly. The main effect of group size was

also significant. However, it is conceivable that there exists a time length shorter than 30 minutes in which the hypothesis would be supported.

This was a surprising result. In addition to previous literature stating smaller groups would outperform large groups, it seemed despite the negative feedback stating frustrations of working in larger groups, the larger groups consistently performed better. This result was further explored in study 2.

Chapter 4:

Study 2 - Effects of Collaboration Tools and Extreme (Small and Large) Group Sizes

Introduction to Collaboration Tools

In study 1, we noted that the large groups consistently outperformed the small groups and that their advantage seemed to keep increasing up to group size 20. This came as a surprise since most of the previous literature discussed in the background supported the finding that the optimal group size was much less than 20. One plausible explanation was that a collaboration tool had inadvertently been introduced that assisted the large groups outperform the small groups; that is, a collaboration tool helped decrease process loss that large groups typically face.

In Woolley and Hashmi (2013), we discussed the types of online tools that can enhance collective intelligence online. Among the plethora of online collaborations tools, it was found, in particular, that multichannel chatrooms and shared online documents help individuals work simultaneously, prevent production blocking, assist equality of contribution and capture ideas/outputs as they occur.

While POGS did not include a version of the multichannel chatroom, it did, have a shared document, the workspace, which all the individuals could view and contribute to simultaneously. The workspace document used a collaboration tool, called Etherpad.

Etherpad was developed to allow multiple people to collaborate synchronously. Similar to Google Docs, users are able to log on virtually and collaborate simultaneously to complete tasks together in real time. Practically, remote users type together in a shared space while being able to 'see' what other users type. Etherpad can be modified to allow a range of flexibility – free flow of text to limited text-specific boxes in a grid. Hence, in the software tools used here, subjects use Etherpad for a variety of tasks ranging all the way from a typing task where subjects recreate a passage to typing numbers in a Sudoku puzzle grid.

In the Wooley et al (2010) study, subjects were invited to be physically present in a room and complete the tasks together. Typically, subjects would 'write down the answers' on the answer sheet we provided to them. One of the subjects, either voluntarily or selected by others, would scribe the answers. It is important to note that while the subjects all worked together, only one was able to write the answer. This is a vastly different dynamic than the online software with Etherpad that we used in Study 1.

Hence, in order to test whether the collaboration tool was a reason why the large groups were outperforming the small groups, it was important to create a condition 'without the collaboration tool'. That is, a condition where the group is not able to type the answers

together. Simulating the Wooley et al (2010) study, this is a condition is where only one person is able to 'scribe' the answers.

The two conditions then became: 1) a condition in which only one person can type in the answers (single scribe) and 2) a condition in which all the people are able to type in the answers (multi-scribe).

Hence, to explore how collective intelligence changes as group size increases with and without the collaboration tool, I set up two conditions (single-scribe and multi-scribe) with two group sizes (4 and 20).

Hypotheses

For this study, three hypotheses were developed.

> *H1: There will be no significant difference in the collective intelligence scores between the small size groups in single-scribe and multi-scribe conditions.*

In a meta-study by Dennis and Williams (2005), the results showed at small group sizes both the electronic brainstorming and verbal brainstorming groups produced about the same number of ideas. This was due to the effect of process losses being at a level where they did not negatively impact the group performance. This is similar to when a group has a collaboration tool and when a group doesn't. For this reason, I believe at the small group

sizes (size 4), a collaboration tool will offer no more benefit to a group than a group without such collaboration abilities. Therefore, groups in single-scribe condition and groups in multi-scribe condition will have similar collective intelligence scores.

H2: At the large group size, groups with multi-scribe will have higher collective intelligence scores than groups with a single-scribe.

Dennis and Williams (2005) found that as group size increased past group size 10, there was a dramatic increase in the number of ideas in the electronic brainstorming groups relative to the verbal brainstorming groups. The meta-study found that process gains of synergy and social facilitation dramatically increased the group performance with the aid of the collaboration tool relative to groups of the same size that did not have the collaboration tool.

We expect to see similar results. At large group sizes (group size 20), the multi-scribe group will have significant advantage and benefit from having a collaboration tool. This means large group sizes in the multi-scribe condition will have significantly higher collective intelligence scores than the large group sizes with single-scribes.

H3: The CI scores of large groups would not be significantly different from the CI scores of the small groups in the single-scribe condition.

Given we tried to simulate face-to-face in the single-scribe condition, the literature found that as group size increases, process losses became more pronounced resulting in decreased group performance (Mueller 2012). Hence, we expect in the single-scribe conditions, the large group sizes will experience process loss at a level that negatively impacts their performance. This will outweigh any benefits that could be derived from having more resources in a large group. As such, large groups will not be significantly different from the small groups in the single-scribe condition.

Method

Data and Setting

Subjects were recruited from MTurk. I kept the same criterion as the first study to qualify a Turker:

- located in the US or Canada
- have successfully completed 1000 or more HITs
- have an approval rate of 95%
- never have completed any of our experiments before

Workers were paid $3.50 for approximately 25 minutes of the experiment. Similar to the first study, all the experiment sessions were run during business hours in the weekday. The batches were released 24 hours in advance. Each session consisted only of groups of 4 or

groups of 20. There were no sessions that had both groups of 4 and 20. Sessions were staggered to run one group size (either 4 or 20) each day of each condition

I had two size groups (group sizes 4 and 20) in two different conditions (single-scribe and multi-scribe) with five groups in each condition:

	Single-scribe	Multi-scribe
Group size 4	5 groups	5 groups
Group size 20	5 groups	5 groups

In the condition of groups in the 'single-scribe' condition, one person was randomly selected in the group who had the ability to scribe the answers in the workspace; where the answers were typed in. Other team members had the ability to 'chat' together. They would not be able to type anything in the workspace. In the alternate condition of 'multi-scribe', all members of the team had the ability to type directly in the workplace. They were also able to chat together. Both conditions had the same set of tasks that were used in the first study (Typing Task, Unscramble words, Matrix reasoning, Sudoku and Brainstorming).

A total of 20 groups were run for this study. Given the need to implement and test new features in POGS, the total time for data collection spanned over five to six months.

Additional Technical Feature of POGS

For this study, an additional technical feature of scribe was introduced. For each session, as Turkers logged in, a Turker would be randomly selected to be the scribe for each group. Once selected to be the scribe, s/he would remain as the scribe through the whole experiment. Neither POGS nor the participants are able to change the scribe once the experiment started.

The scribe and other participants would know who the scribe was both by stating this in the introduction and directions as well as displaying the scribe visually throughout the session (as shown in Figure 5) in the chatroom. Figure 5 shows that 'Nitro' is the scribe while the other participants are not.

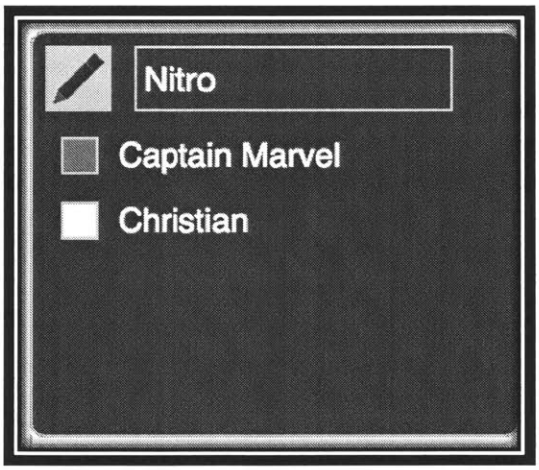


Figure 5: The visual identification of who the scribe is

During pilot tests, we found a glaring issue inherent in POGS. Many times Turkers would simply login to POGS but not actively participate; that is, once logged in, the Turker would

neither chat nor contribute to the experiment in any other way. This issue was a serious problem if that particular Turker was randomly selected as the scribe. In which case, the whole session was wasted as no answers were recorded. Turkers were able to do this as POGS allowed Turkers to login and remain waiting in a 'waitroom' with a countdown until the experiment were to start; sometimes as early as 24 hours in advance. At the start time of the experiment, POGS would then automatically advance the Turkers through the experiment. They would appear in the final list of 'present' Turkers and be paid the full amount on MTurk. Hence, Turkers easily discovered this exploit; that they only had to login but perhaps not actually be present when the experiment started. In an essence, we had simulated the Ringelmann effect virtually.

After discussions on the different MTurk discussion groups, I discovered dedicated Turkers who had posted this exploit. As our task was one of the higher paid HITS, it naturally attracted more attention. In addition to asking them to take down their posts, I engaged them to understand how to prevent this from happening. As such, we then introduced an additional check at the start of the experiment – a screen that asked each participant to confirm attendance by checking a box as soon as the experiment would actually start. They would have 15 seconds to confirm participation. If they did not confirm, then POGS would not advance them.

After the confirmation feature was added, a scribe would only be selected from amongst the active participants. We decided to have the scribe feature be 'voluntary' as opposed to an

'obligation'. If this were to simulate a face to face group, the subject always had the option to refuse to be the scribe and hence we would turn them away from that study as well as not pay them. Similarly, we decided if a Turker was randomly selected to be a scribe, the Turker was given the choice to accept. If the Turker did not accept, they would no longer participate in the experiment and hence not be paid. However, they were still eligible to participate in future studies.

Dependent Variable

Similar to study 1, the dependent variable is the collective intelligence score.

Independent Variables

In this particular study, my set of independent variables was the group size and the two conditions of single-scribe and multi-scribe. I had two group sizes: group size 4 and group size 20. I had two conditions: single-scribe and multi-scribe.

Results

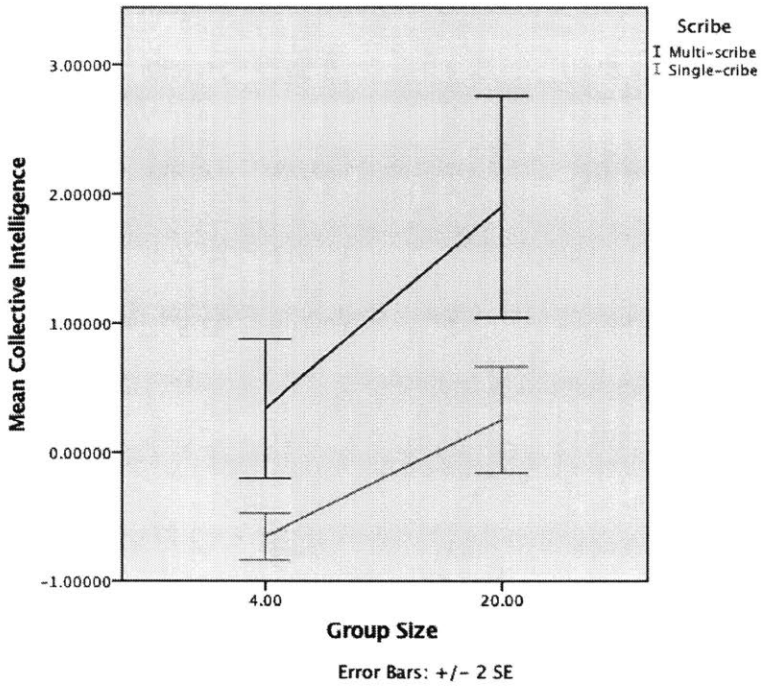
At a first glance, I calculated the collective intelligence scores for each group size at each condition. Table 4 below shows the descriptive statistics. I had a total of 22 groups. Group size 20 in the multi-scribe condition had highest mean for collective intelligence score (1.90 score) while the group size 4 in the single-scribe condition had the lowest mean (-0.65).

Table 4: Descriptive Statistics of Scribe Study

		Collective Intelligence		
Scribe	Group Size	Mean	Standard Deviation	Total N
Multi-scribe	4.00	.34033	.60219	5
	20.00	1.90109	.95683	5
Single-scribe	4.00	-.65145	.38410	5
	20.00	.25691	.54484	7

A visual graph confirms the groups in multi-scribe condition outperformed the groups in the single-scribe condition. Graph 2 further shows that within each group size, the groups in the multi-scribe condition outperformed the groups in the single-scribe.

Graph 2: Results of effects of Scribe Condition and Group Size



In order to understand the interaction effects, I ran a two-way Anova which resulted in showing the scribe conditions and group size each differed significantly, F(1,9)=38.385, p=0.000 and F(1, 9)=33.592, p=0.000 respectively (Table 5). Furthermore, there is no statistically significant interaction between scribe and group size for the Collective Intelligence score, $F(1, 31) = 2.345$, $p = .136$, , partial $\eta^2 = .070$.

Table 5: Tests of Between-Subjects Effects

Dependent Variable: Collective Intelligence

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared

Corrected Model	26.704[a]	3	8.901	29.350	.000	.740
Intercept	5.700	1	5.700	18.794	.000	.377
Scribe	11.611	1	11.611	38.285	.000	.553
Group_size	10.188	1	10.188	33.592	.000	.520
Scribe * Group_size	.711	1	.711	2.345	.136	.070
Error	9.402	31	.303			
Total	36.153	35				
Corrected Total	36.106	34				

a. R Squared = .740 (Adjusted R Squared = .714)

Conducting a test of contrasts from the two-way Anova to understand simple main effects, I found there was a statistically significant difference in mean Collective Intelligence score between groups in the single-scribe and multi-scribe conditions of size 4, $F(1, 31) = 12.691$, $p = .001$, partial $\eta^2 = .290$ and size 20 $F(1, 31) = 25.998$, $p = .000$, partial $\eta^2 = .456$. (Table 6). Each F tests the simple effects of Scribe within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Table 6: Test of Contrasts

Dependent Variable: Collective Intelligence							
Group Size		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
4.00	Contrast	3.849	1	3.849	12.691	.001	.290
	Error	9.402	31	.303			
20.00	Contrast	7.885	1	7.885	25.998	.000	.456
	Error	9.402	31	.303			

Given that the particular simple main effect is statistically significant, I considered the difference in mean collective intelligence score between groups in the single-scribe and multi-scribe conditions of both different group sizes. I find that in group size 20, the mean collective intelligence score was 1.644 (95% CI, 0.987 to 2.302) points higher for groups in the multi-scribe condition. Similarly, for groups of size 4, I find that the mean collective intelligence score was .992 (95% CI, .424 to 1.560) points higher for groups in the multi-scribe condition (Table 7).

Table 7: Pairwise Comparisons

Dependent Variable: Collective Intelligence

Group Size	Mean Difference (Multi-scribe – Single Scribe)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
4.00	.992*	.278	.001	.424	1.560
20.00	1.644*	.322	.000	.987	2.302

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Bonferroni.

When comparing within the groups in the single-scribe and multi-scribe conditions, I find that, in the multi-scribe groups, the mean collective intelligence score for groups of 4 was 1.561 (95% CI, -2.271 to -0.850) below the mean collective intelligence score for groups of 20. In groups with single-scribes, groups of 4 mean collective intelligence score was 0.908 (95% CI, -1.409 to -0.408) below the mean scores of groups of 20 (Table 8).

Table 8: Pairwise Comparisons

Dependent Variable: Collective Intelligence					
Scribe	Mean Difference (Group size 4 – Group size 20)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound

Multi-scribe	-1.561*	.348	.000	-2.271	-.850
Single-scribe	-.908*	.245	.001	-1.409	-.408

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Bonferroni.

Discussion

H1: There will be no significant difference in the collective intelligence scores between the small size groups in single-scribe and multi-scribe conditions.

Unsupported. Interestingly enough, even at the small group size, there was a statistical difference between the single-scribe and multi-scribe groups.

H2: At the large group size, groups with multi-scribe will have higher collective intelligence scores than groups with a single-scribe.

Supported. Our study did show that at group size 20, the multi-scribe groups significantly outperformed the single-scribes.

H3: The CI scores of large groups would not be significantly different from the CI scores of the small groups in the single-scribe condition.

Unsupported. Within the single-scribe condition, the large groups significantly outperformed the small groups.

Our main effect of the scribe condition (multi-scribe and single-scribe) was significant. That implies, as we had hypothesized, allowing groups to collaborate synchronously does in fact significantly improve their CI score. Hence, a collaboration tool, *Etherpad,* was playing a role in improving performance of the large group sizes. However, a surprising result was that the multi-scribe groups outperformed the single-scribe groups at both the small and large group sizes. Furthermore, we found that, even within the single-scribe groups, the large groups significantly outperformed the small groups.

One other plausible explanation was that there is a large interval between the small and large group size. In smaller intervals, we might find the performance we hypothesized. This was explored in the third study.

Chapter 5:

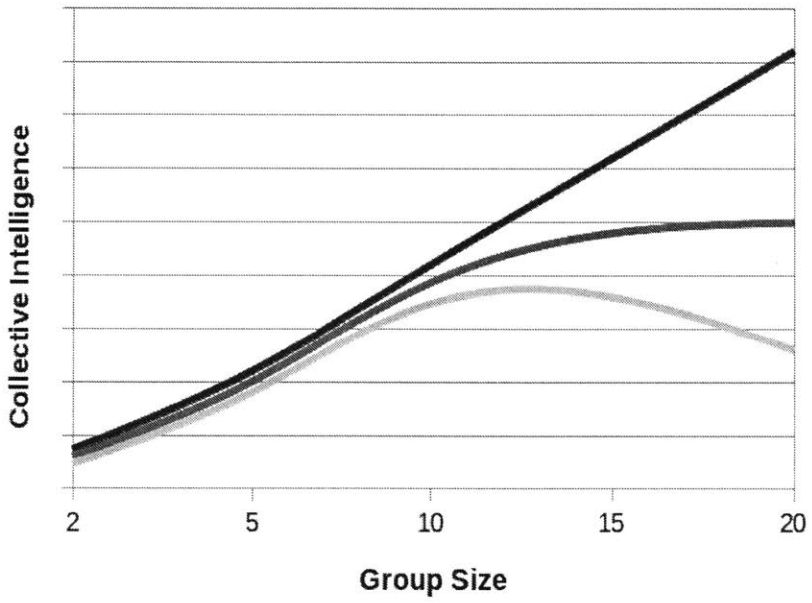
Study 3 - Effects of a Range of Group Sizes

Introduction to Effects of Group Sizes and Optimal Group Size

One of the interesting results from the second study was that the larger groups outperformed the smaller groups in both the single-scribe and multi-scribe conditions. Given that most of the literature has shown smaller group sizes to outperform larger group sizes with the optimal group size being between 5 and 7; this finding contradicting that and it warranted a study that explored collective intelligence over a range of group sizes.

Specifically, we were curious about how the CI score might be changing in group sizes between 4 and 20. Is it increasing steadily? Does it peak at a number in between and then plateau? Or does it peak at a number closer to group size 20 and then begin to decrease? Furthermore, if it is increasing steadily up to size 20, at what point do the process losses begin to negatively impact the performance of the group? Essentially, at what point does the CI score stop increasing beyond group size 20. Graph 3 below shows the three different trajectories that could potentially occur for group sizes between 4 and 20.

Graph 3: Three different trajectories of CI as group size increases



In addition, we saw the main effects of single-scribe and multi-scribe being significant but the interaction effect was not significant in study 2. It is possible that with shorter intervals of group size, an interaction effect of group size and scribe might become pronounced and hence, the results would show different trajectories for single-scribe and multi-scribe.

Finally, can we find an 'optimal group size' in both of the conditions? For the purpose of this thesis, 'optimal group size' is the size at which the highest CI Score is obtained.

These questions motivated the third study where we studied the effects of a range of group sizes in both the single-scribe and multi-scribe conditions of group sizes 5, 10, 15, 20, 25, 30, 35 and 40.

Hypothesis

For this study, we developed three hypotheses:

> *H1: To the degree that single-scribe reflects face to face groups, optimal group size would be between 5 and 10.*

As the single-scribe condition was meant to simulate a face to face group, most literature referred to the optimal group sizes ranging between 5 and 7. Hence, it is highly possible the optimal group size would be between 5 and 10.

> *H2: The optimal group size for multi-scribe would be greater than 20.*

The Mao et al study (2013) showed groups of size 32 with the best performance when compared to group sizes 1, 2, 4, 8 and 16. In addition, that study also recruited Turkers and the complete study was conducted online. Hence, there is evidence to believe the optimal group size could be greater than 20 for online groups using collaboration tools like the Etherpad editor that was used in the multi-scribe condition .

H3: In both the conditions (single-scribe and multi-scribe), as group sizes increases, at some point the CI score will start to decrease; that is, a curvilinear relationship (inverted-U) exists.

Most of the literature on group size discusses how group performance, decision making, effectiveness all 'decrease' as the group size increases after the optimal group size is reached (Mueller 2012, Hackman 2002, Blenk et al. 2010). None of the literature discusses or states a result that 'plateaus'. Hence, I believe, after the optimal group is reached, the CI will start to decrease as process losses will become more pronounced and their negative impact will outweigh the benefits of having more resources in a larger group.

Method

Determining Group Sizes

While Mao et al (2013) had selected groups with an even number of members that doubled in each interval, research on group size has shown that there is evidence of a difference in odd or even group sizes (Menon and Phillips 2008). Furthermore, the strategy to double at each interval potentially created large gaps in between group sizes where CI could potentially change. Finally, the successful tested technical upper limit of a group size that could complete the experiment in POGS was 40.

As such, the following group sizes were selected for this study: 5, 10, 15, 20, 25, 30, 35 and 40 in each of the two conditions (single-scribe and multi-scribe). This set of group sizes

included both even and odd groups at relatively small intervals and reached the maximum capacity of POGS.

Determining Sample Size

I conducted a power analysis for a two-way ANOVA for 2x8 groups in G*Power to determine a sufficient sample size using an alpha of 0.05, a power of 0.80, degrees of freedom at 7 and a large effect size (f = 0.41) calculated using our pilot study data. Based on the aforementioned assumptions, the desired minimum sample size was 104. At that minimum, it translates to 6 groups per group sizes. However, that is the minimum. Ultimately, for each group size, I collected data from at least 10 groups. Altogether, I aimed to collect data from 160 groups which include a total of 6200 Turkers.

Subject Recruitment

Subjects were recruited randomly from Amazon. Given the high number of Turkers that was required for this study, I reduced the success and number of HITs accepted criterion slightly:

- Have an 80% or more approval rate
- Have successfully completed 500+ hits accepted
- Located in the US or Canada
- Never have completed any of our experiments before

Unless it was a holiday or a special day (natural disaster, etc), four sessions were run throughout the week, including weekends, at 12:00pm, 3:00pm, 6:00pm, 9:00pm (Eastern Time) daily. Of the four sessions, each session was randomly assigned the single-scribe or multi-scribe condition; similar to flipping a coin. Up to 60 HITs were made available in each batch; that is, with the expectation that up to 60 Turkers could log in and complete the study at each session. Depending on the actual number of Turkers who logged in, the Turkers were randomly assigned to a group size.

Turkers were paid $4.00 for approximately 25 minutes of the session. Not including the pilot tests, this study took seven weeks to complete.

Subject Assignment to Group Sizes

I developed a random matrix generator (RMG) in the Python computer programming language to randomly select and vary group sizes in each session according to the number of subjects from MTurk who actually log in. As input, I gave the RMG the largest number of subjects that I can expect to show up and the array of group sizes pertinent to my study. The RMG then generates a table of combinations for all the possible number of subjects who could show up (1 through the largest number) using the study group sizes of interest.

For example, in this third study, we were interested in group sizes 5,10,15,20,25,30,35 and 40. As I intended to over-recruit, I had 60 HITS for each batch; that is, a total of 60 subjects could potentially log in at the same time. Hence, the table that the RMG developed accounted for the possibility of any number of subjects between 1 and 60 subjects up. There is a high probability to expect the total number of subjects to not be evenly distributed into all of our group sizes. That is, if 33 subjects logging in; while 30 is a group size, there will be three subjects left over which is not a group size we are interested in. Hence, I needed to compensate for the overflow subjects. One option was to reject them. Another option was to simply create groups with the overflow subjects. Given I would be compensating them financially anyways and POGS does not handle overflows automatically, I developed RMG to add additional group sizes in order to create groups with the overflow subjects. Thus, the final array of group sizes became 1, 2, 3, 5, 10, 15, 20, 25, 30, 35 and 40.

Given the final array of group sizes, the RMG iterated through all the possible combinations of group sizes and developed a matrix of all the possible combinations in which the group sizes only appear once iterating through all the possible numbers of Turkers logging in (that is, 1 through 60). The advantage of a group sizes appearing only once was to prevent the possible bias of multiple sessions of the same group size being collected at a certain time. With each group size appearing only once, that ensured a random distribution of when the data for that group size is collected. For example, in the case when 10 subjects showed up, RMG would produce the following matrix:

10,10

10,5,3,2

The "10" at the beginning of each row means that there are 10 subjects available to be assigned to groups. The first row indicates that all 10 subjects are assigned to the same group. The second row indicates that the 10 subjects are assigned to three groups of sizes 5, 3, and 2.

From this matrix, the system will then randomly select one row to use for each session. As each group size gets completed (that is, we are able to have 10 groups for that group size), RMG can then be switched to creating and selecting combinations from the remaining group sizes.

Here is an example of what the RMG produces, where the first number represents the number of Turkers who have logged in followed by the group sizes in which to distribute the Turkers:

60,25,15,10,5,3,2	*40,25,10,3,2*	*20,10,5,3,2*
59,25,20,10,3,1	*39,20,15,3,1*	*19,15,3,1*
58,30,25,3	*38,35,2,1*	*18,10,5,2,1*
57,40,10,5,2	*37,20,15,2*	*17,10,5,2*
56,40,10,3,2,1	*36,25,5,3,2,1*	*16,10,3,2,1*

55,35,15,3,2	*35,20,10,3,2*	*15,10,5*
54,30,15,5,3,1	*34,20,10,3,1*	*14,10,3,1*
53,25,15,10,3	*33,30,3*	*13,10,3*
52,25,15,10,2	*32,25,5,2*	*12,10,2*
51,30,10,5,3,2,1	*31,15,10,3,2,1*	*11,5,3,2,1*
50,30,10,5,3,2	*30,20,5,3,2*	*10,10*
49,20,15,10,3,1	*29,20,5,3,1*	*9,5,3,1*
48,25,15,5,2,1	*28,20,5,3*	*8,5,3*
47,25,15,5,2	*27,20,5,2*	*7,5,2*
46,35,5,3,2,1	*26,20,3,2,1*	*6,5,1*
45,30,10,5	*25,15,5,3,2*	*5,3,2*
44,40,3,1	*24,20,3,1*	*4,3,1*
43,40,3	*23,15,5,2,1*	*3,3*
42,30,10,2	*22,20,2*	*2,2*
41,25,10,3,2,1	*21,15,5,1*	*1,1*

This table is fed to the POGS server and depending on how many Turkers actually showed up, the group sizes were assigned accordingly. For instance, if 21 Turkers were to login for the session, POGS would refer to the matrix produced by RMG, locate the row with 21 participants and then create three groups of 15, 5 and 1 person. These three groups would run for that session.

Hence, through this method, random collection of the data for each group size was ensured.

Additional Technical Features of POGS

For this study, we increased POGS capacity from having an upper limit of 20 subjects in one group to an upper limit of 40 subjects in one group. Also, as a result of needing to vary group sizes in each session, an additional feature was introduced that allowed the researcher to provide the matrix and POGS would then refer to the table and create groups accordingly.

Dependent Variable

Similar to the previous study, the dependent variable was the collective intelligence score.

Independent Variables

In this particular study, the set of independent variables was the group size and the scribe condition. There were a total of 8 group sizes: 5, 10, 15, 20, 25, 30, 35 and 40 and two scribe conditions: single-scribe and multi-scribe.

Results

I calculated the CI Score for each group size in each condition. Table 9 shows the descriptive statistics. Altogether, there were 179 groups (87 in the multi-scribe condition and 92 in the single-scribe condition). This is the final number of groups that I used for my analysis; I did not include any group that had any technical errors or where other issues arose (e.g. a Turker who made racist comments and upset other Turkers).

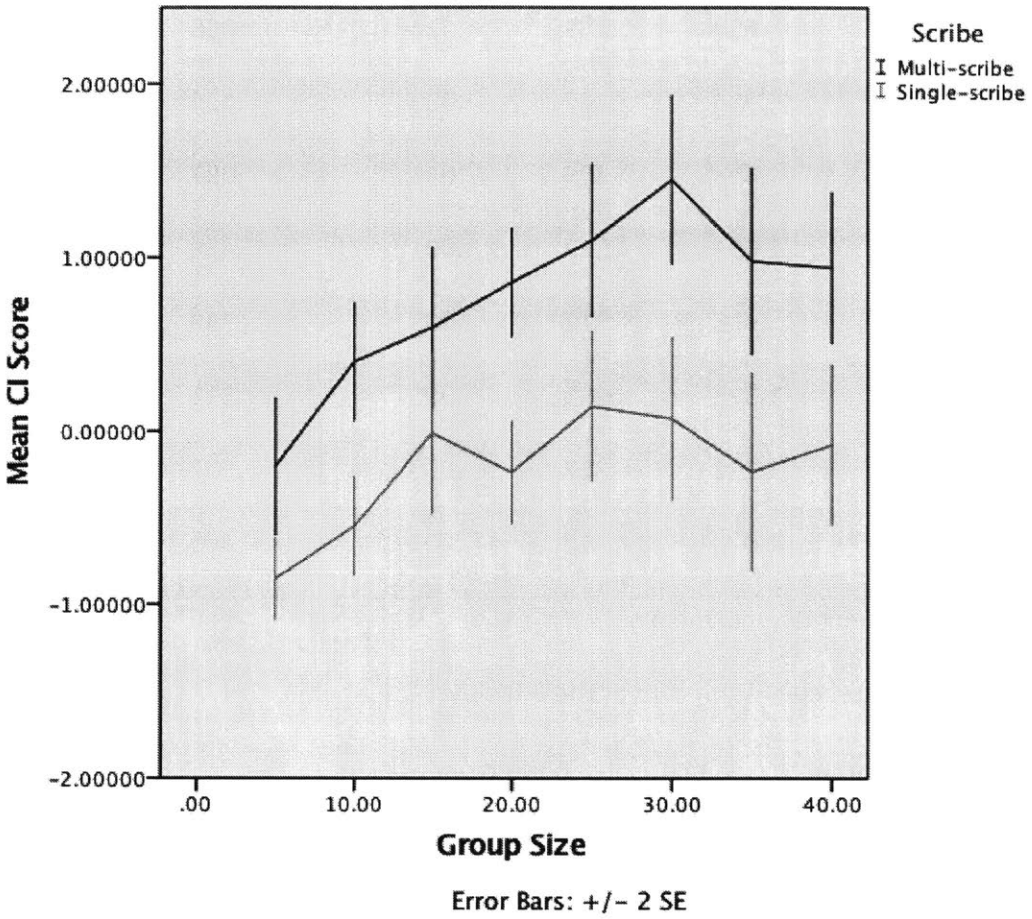
Table 9: Descriptive Statistics of Effect of Group Size Study

Scribe	Group Size	CI Score Mean	CI Score Std. Deviation	Total N
Multi-scribe	5.00	-.2009816	.63002944	10
	10.00	.4040024	.68332128	16
	15.00	.6005581	.77046264	11
	20.00	.8604787	.49769762	10
	25.00	1.1016054	.70044632	10
	30.00	1.4501238	.77070683	10
	35.00	.9819854	.85277441	10
	40.00	.9452762	.69220330	10
	Total	.3977319	1.02002876	87
Single-scribe	5.00	-.8505378	.46969779	15
	10.00	-.5457656	.49866353	12
	15.00	-.0075504	.73767478	10
	20.00	-.2332889	.47053384	10

25.00	.1468018	.68975815	10
30.00	.0769712	.91898753	15
35.00	-.2302549	.90876854	10
40.00	-.0751521	.73811153	10
Total	-.4084814	.79700179	92

A visual graph, Graph 4, confirms the multi-scribe groups outperformed the groups with single-scribes across all group sizes.

Graph 4: Results of the effects of Scribe Condition and Group Size



A pairwise comparison in each group size between each condition was conducted to assess whether the differences were significant (Table 10)

Table 10: Pairwise Comparisons

Dependent Variable: CI Score

	Mean Difference (Multi-scribe –			95% Confidence Interval for Difference[b]	
Group Size	Single-scribe)	Std. Error	Sig.[b]	Lower Bound	Upper Bound

5.00	.650*	.285	.024	.088	1.212
10.00	.950*	.267	.000	.424	1.475
15.00	.608*	.305	.048	.007	1.210
20.00	1.094*	.312	.001	.478	1.709
25.00	.955*	.312	.003	.339	1.570
30.00	1.373*	.285	.000	.811	1.935
35.00	1.212*	.312	.000	.597	1.828
40.00	1.020*	.312	.001	.405	1.636

Based on estimated marginal means

*. The mean difference is significant at the 0.01 level

b. Adjustment for multiple comparisons: Bonferroni.

This indicated that at each group size, there was a significant difference between both the conditions.

Similar to the second study, we were interested to see the interaction effects of scribe and group size. Furthermore, it was important to understand whether there was a significant curvilinear relationship between scribe and group size. Given group size was a positive integer, in order to avoid multicollinearity, I mean centered group size before squaring it and creating the quadratic term. As such, I ran two regression models one with the linear interaction effect and one that included the quadratic interaction effect.

Hence, our first model became:

$$CI\_Score = \beta_1 + \beta_2 * GS + \beta_3 * S + \beta_4 * (GS * S)$$

Our second model became:

$$CI\_Score = \beta_1 + \beta_2 * GS + \beta_3 * S + \beta_4 * (GS * S) + \beta_5 * GS^2 + \beta_6 * (GS^2 * S)$$

Where:

GS = Group Size

S = Scribe Condition

GS^2 = Mean Centered Group Size Squared

The results showed our first model where scribe, group size and their linear interaction effect significantly predicted the CI score, $F(3,175) = 38.408$, $p<.05$. However, the interaction effect was not significant.

Our second model also significantly predicted the CI Score, $F(5, 173) = 28.972$, $p<0.01$. While the quadratic interaction term was not significant, the R square change in the second model compared to the first model was in fact significant.

After running the two models, the result of the quadratic term (GS^2) was found to be significant in both the models. This along with the visual graph, confirms a curvilinear (inverted-U) relationship. However, the quadratic interaction term ($GS^2 * S$) is not significant.

This means for both the scribe conditions (single-scribe and multi-scribe), we cannot conclude the inverted-U relationships are not parallel. Table 11 summarizes the results for both of the models.

Table 11: Linear Regression Models

	Model 1			Model 2		
Variable	*B*	Std. Error	Beta	*B*	Std. Error	Beta
(Constant)	0.044	0.166		-0.085	0.163	
Scribe	-0.741	0.23	-0.401**	-0.652	0.224	-0.353**
Group Size	0.032	0.007	0.405**	0.053	0.009	0.668**
GroupSize*Scribe	-0.011	0.009	-0.168	-0.018	0.012	-0.274
Squared Mean Centered Group Size				-0.002	0.001	-0.363**
Scribe * Squared Mean Centered Group Size				0.001	0.001	0.073
R Square		0.397			0.454	

F for change in R Square	38.408*	28.972**

*p<0.05 **p<0.01

As the interaction effect was not significant, two different approaches were then used to estimate optimal group sizes. The first approach was simply to observe that the highest means in each condition were as follows (see Table 9 descriptive table):

- Single-scribe: at group size 25 with a CI Score mean of .147
- Multi-scribe: at group size 30 with a CI Score mean of 1.450

The second approach was to create models which were then used to estimate optimal group sizes. To do this, I first removed the insignificant terms such that our model became:

$$CI\_Score = \beta_1 + \beta_2 * GS + \beta_3 * S + \beta_4 * GS^2$$

Where:

GS = Group Size

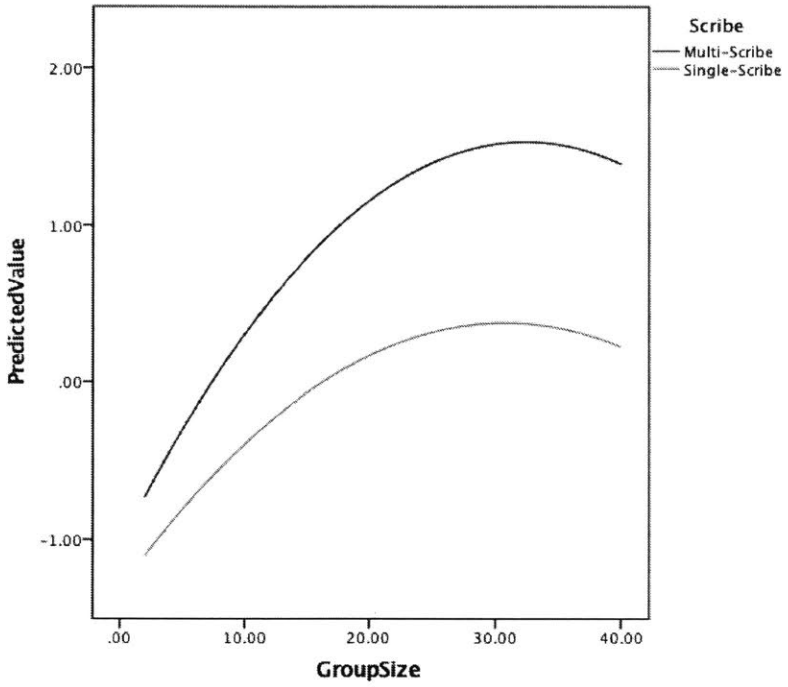
S = Scribe Condition

GS^2 = Mean Centered Group Size Squared

Using this model, I predicted CI scores at smaller intervals (intervals of 1). I found the peak for multi-scribe to be at 34 and for the single-scribe condition, group sizes 29, 30 and 31 all

had similar values and the highest CI scores (Graph 6). However, these are predicted values, not actual peak values.

Graph 6: Predicted Values of CI Score



Discussion

There were three hypotheses formulated.

H1: To the degree that single-scribe reflects face to face groups, optimal group size would be between 5 and 10.

Unsupported. In our data set, we found the optimal group size for single-scribe to be at 25. This is higher than what we hypothesized, and is a surprising result in light of most of the prior literature.

H2: The optimal group size for multi-scribe would be greater than 20.

Supported. In our data set, we found the optimal group size for multi-scribe to be at 30.

H3: In both the conditions (single-scribe and multi-scribe), as group sizes increases, at some point the CI score will start to decrease; that is, a curvilinear relationship (inverted-U) exists.

Supported. Our data set shows the quadratic term (GS^2) to have been significant and along with the visual graph, this supports that a curvilinear relationship (inverted-U) exists. That is, the CI score increases as group size increases but at some point, the CI score begins to decrease as the group size continues to increase.

There are many plausible explanations of why we saw significantly larger than expected optimal group size for the single-scribe. One might be that the single-scribe does not accurately reflect a face-to-face group. While having one person scribe is reflective of the process within a face-to-face group, a face-to-face communication mechanism is inherently different than an online chat interface. Furthermore, in addition to being able to verbally

communicate, face-to-face groups are also able to use body language. And it is much more difficult to have 20 or 40 people communicating together in a face-to-face group as there may be people talking over each other, shouting across to others and even not paying attention. The online version potentially stream lines the communication process in the chatroom where each line represents one person communicating in the order that the person types. Furthermore, there is no concept of 'distance' in the chat while distance from one person to another in a large group might play additional roles in the communication process of face-to-face groups. Communication was also seen to be a crucial factor that explained the CI score in the Wooley et el (2010) study. Finally, there may be other hidden nuances in face-to-face groups of which we only captured one aspect (single-scribe) in our experiments.

Another plausible explanation might be that the tasks selected for the CI battery tasks get greater benefit from large groups. Steiner's Taxonomy (1972) categorizes tasks into the following:

- Additive: Individual inputs are added together (e.g. shoveling snow)
- Compensatory: Decision is made by averaging together individual decisions (e.g. guess weight of a person)
- Disjunctive: Group selects one solution or product from a pool of members' solutions or products (e.g. math problem)

- Conjunctive: All group members must contribute to the product for it to be completed (e.g. climb a mountain together tied with ropes)
- Discretionary: Group decides how individual inputs relate to group product (e.g. could vote for the best answer or leader decides)

For the five tasks included in the CI battery of tasks, they could be categorized as follows:

- Typing: Additive
- Brainstorming: Additive
- Sudoku: Discretionary
- Matrix Solving: Disjunctive
- Unscrambling: Disjunctive

In additive tasks, as group size increases so does the performance. Kerr and Bruun (1983) found that as group size increases, performance in disjunctive tasks also increases while performance decreases for conjunctive tasks. Their experiments involved face-to-face groups of sizes 2, 4, and 8. Hence it is theoretically possible, that the tasks in the CI batter of tasks were advantageous to large groups.

Another plausible explanation is the embedded collaboration and coordination tools that are present when conducting a task virtually or on the Internet. Similar to the results seen in the Mao et al (2016) study, the presence of the Internet may be facilitating the collaboration and

coordination that benefits groups in both the single-scribe and multi-scribe conditions. The subjects are already familiar with working remotely on computers and using online virtual collaboration tools. POGS is embedded with a multitude of collaboration tools and a layout aimed to ease and facilitate group work. This is vastly different than any equivalent pen and paper task. As such, this may also explain why the results have consistently shown higher optimal group sizes than what is found in the majority of literature.

As future steps, an in-depth analysis at the task level should be conducted to further understand which tasks were beneficial to the different conditions in the different group sizes. In addition, future studies can be designed to include more interdependent tasks that span across the different task types.

Chapter 6: Conclusion

This thesis set out to explore the effects of group size on collective intelligence. In order to run the studies, the Platform for Online Groups (POGS) was modified to increase the capacity, run different group sizes within a session, and add scribe features. Amazon Mechanical Turk was used to recruit subjects for each of the studies.

In the first study, we explored the effects of time pressure and group size. Five groups of size 4 and size 20 were placed in each of the four different time lengths Tx1 (30 minutes), Tx1.5 (45 minutes), Tx2 (1 hour) and Tx3 (1.5 hours). We found, in general, the large groups outperformed the small groups in all four time lengths.

This came as a surprise as we expected small groups to outperform at the shorter time lengths. This prompted us to explore whether this was because we had inadvertently introduced a collaboration tool, called Etherpad. Hence, we tested out our theory by creating a non-collaborative scenario where only one person had the ability to type in the answer, the 'single-scribe' condition. The alternative condition was where everyone in the group had the ability to type in the answer, the 'multi-scribe' condition.

The two scribe conditions (single-scribe and multi-scribe) were used to test whether the collaboration tool was in fact benefiting the large groups. Five groups of sizes 4 and size 20

were placed in both of the scribe conditions in Tx1 (30 minutes) time length. The results showed that multi-scribe outperformed single-scribe in both the small and large group sizes. Furthermore, the large groups outperformed the small groups in both scribe conditions. Our main effects of scribe and group size were also significant. However, the interaction effect was not significant.

From this study, we learned and confirmed that the collaboration tool (as in the multi-scribe condition) was benefiting the groups at both the small and large group sizes. The interaction effect not being significant came as a surprise. This implied that it was not just the collaboration tool benefiting the large groups. However, given the study only used two group sizes, it was important to explore whether this held true with a much larger study that involved a range of group sizes.

Hence, our third study set out to explore a range of group sizes in the two scribe conditions. There were eight different group sizes selected: 5, 10, 15, 20, 25, 30, 35 and 40. Data from a minimum of 10 groups for each of the different groups in both of the scribe conditions was collected and analyzed. From this, the results showed that a curvilinear (inverted-U) relationship existed between the CI score and group size but the interaction term was not significant. As the group size increased, the CI score increased until at some point it started to decrease. Furthermore, in our data set we discovered the optimal group sizes for single-scribe (25) and for multi-scribe (30). The single-scribe optimal group size came as a surprise as it was much larger than what the previous literature found.

There were many potential explanations of why we saw a larger optimal group size for the single-scribe conditions that included whether the single-scribe truly reflected a face-to-face group and if the tasks inherently benefited large groups. Both of these factors need to be further explored in future experiments.

In sum, the first study showed that large groups consistently outperformed the small groups. The second study results showed that a collaboration tool did in fact contribute to increasing the CI score for groups. The third study showed that there exists a curvilinear relationship of CI and group size. Both the studies showed the CI score peaking at a much higher size than what has been found in previous literature.

And while the studies display a consistent result and built upon one another, none of the studies have been able to explain why the CI scores peak at larger group sizes than known in previous literature. As next steps, it is important to understand the underlying mechanism of why this is the case. For instance, as a first step, a deeper analysis of the results at the task level could be conducted to assess whether the tasks biased the results. If so, other tasks could also be included in the CI battery of tasks.

Furthermore, while it is positive to see a larger optimal group size for both single-scribe and multi-scribe, these studies were conducted in laboratory settings. It is important to conduct

empirical studies and assess the effect on businesses; especially as the world moves towards more technologically advanced interactions, etc.

Finally, the optimal group size as seen in this data set is the group size that achieves the highest CI score value. It is also important to understand and value the gain of each additional group member in order to achieve the higher CI score. It is conceivable in a scenario where each additional gain in the CI score is very valuable to the team (e.g. a difference of millions of dollars) and hence any marginal gain would be of great value. Conversely, it is also possible that each marginal gain in the CI score is very expensive and of lesser value (e.g. difference of a few hundred dollars). In either of the scenarios, knowing the marginal value of each additional member and the change in the CI score would be important. Should businesses employ a large team sizes to gain maximum benefit or save the additional costs of more people but settle for suboptimal performance?

While a longitudinal empirical study with large group sizes with real teams would build more confidence, one other implication is for businesses that regularly engage geographically dispersed employees in teams. These studies show large geographically dispersed teams have high CI scores. Hence, businesses should feel more comfortable and confident in the employment of remote workers who may be geographically dispersed or simply work from home.

Furthermore, independent of what group size should businesses employ, one conclusion is for certain across businesses. The use of online tools, internet, virtual environments should be emphasized drastically. Even if the groups within the business are all local and face-to-face, our studies, that inherently have collaboration and coordination tools embedded, show a drastic higher performance at larger group sizes than what the previous literature shows. Business should consider rethinking and modeling around technology that, even at the local face-to-face level, embraces and enhances group work.

These questions, ideas and more need to be explored to fully understand the effects of group size on collective intelligence. This thesis sets the path towards that.

References

Aggrawal, I., & Woolley A.W. (2013) Do you see what I see? The effect of members' cognitive styles on team processes and performance. Organizational Behavior and Human Decision Processes 122:92-99.

Anderson, L., & Frank, J. (1971). Effects of task and group size upon productivity and member satisfaction. Sociometry, 34, 135-149

Baron-Cohen S., Wheelwright S., Hill J., Raste Y. & Plumb I., (2001) The 'Reading the mind in the eyes' test revised version: A study with normal adults, and adults with Asperger Syndrome or High-Functioning autism. J. Child Psychol. Psychiatry 42, 241.

Bell, Suzanne T. (2007) "Deep-level composition variables as predictors of team performance: A meta-analysis." Journal of Applied Psychology, Vol 92(3), May 2007, 595-615

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

Berkowitz, M. I. (1958). An experimental study of the relationship between group size and social organization. Doctoral Dissertation. Yale University. New Haven, CT.

Blenko, M.W., Rogers, P., Mankins, M.C., (2010) Decide and Deliver: Five Steps to Breakthrough Performance in Your Organization, Harvard Business Review

Bottger, P.C., Yetton, P.W., (1987) Improving group performance by training in individual problem solving.

Brandon D. M., Long J. H., Loraas T. M., Mueller-Phillips J, & Vansant B. (2014) Online Instrument Delivery and Participant Recruitment Services: Emerging Opportunities for Behavioral Accounting Research. Behavioral Research in Accounting: Spring, Vol. 26, No. 1, pp. 1-23.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. Perspectives on Psychological Science, 6(1), 3-5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behavior Research Methods, 46, 112–130.

Davis, J. H., (1980) Group Decision and Procedural Justice. In M. Fishbein (Ed.), Progress in Social Psychology (Vol. 1), Erlbaum, Hillsdale, NJ.

De Dreu, C.K.W., (2003). Time pressure and closing of the mind in negotiation. Organizational Behavior and Human Decision Processes 91, 280–295.

De Grada, E., Kruglanski, A.W., Mannetti, L., Pierro, A., (1999). Motivated cognition and group interaction: need for closure affects the contents and processes of collective negotiation. Journal of Experimental Social Psychology 35, 346–365.

Deary I. J., (2000) Looking Down on Human Intelligence: From Psychometrics to the Brain. Oxford Univ. Press, New York

Deary, I. J., (2000) Looking down on human intelligence: From psychometrics to the brain. Oxford University Press, New York

Dennis, A.R., (1996). Information exchange and use in small group decision making. Small Group Research 27, 532–550.

Dennis, Alan & L. Williams, Michael. (2005). A Meta-Analysis of Group Side Effects in Electronic Brainstorming: More Heads are Better than One.. IJeC. 1. 24-42. 10.4018/978-1-59904-393-7.ch013.

Difallah, D. E., et al. (2015). The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, International World Wide Web Conferences Steering Committee: 238-247.

Driskell, J. E., Salas, E., & Johnston, J. (1995). Is stress training generalizable to novel settings? Paper presented at the 10th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Durham, C.C., Locke, E.A., Poon, J.M.L., McLeod, P.L., (2000). Effects of group goals and time pressure on group efficacy, information-seeking strategy, and performance. Human Performance 13, 115–138.

Edmondson A, (1999) Psychological safety and learning behavior in work teams. Administrative Science Quarterly 44:350–383.

Engel, D., Woolley A.W., Jing L.X., Chabris C.F, Malone T.W., (2014) Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face

Engel, D., Woolley, A. W., Aggarwal, I., Chabris, C. F., Takahashi, M., Nemoto, K., Kaiser, C., Kim, Y. J., & Malone, T. W. (2015). Collective intelligence in computer-mediated collaboration

emerges in different contexts and cultures. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2015), Seoul, Korea.

Fink, C., & Thomas, E. (1963). Effects of group size. Psychological Bulletin, 60, 371-384

Fox, D., Herrold, K., Lorge, I., & Weltz, P. (1953).Twenty questions: Efficiency in problem solving as a function of size of group. Journal of Experimental Psychology, 44, 360-368

Gallupe, R., Dennis, A.R,, Cooper, W,H, Valacich, J. S., Bastianutti, L. M., & Nunamaker, J. F. Jr. (1992), Electronic Brainstorming and Group Size. Academy of Management Journal, 35, 350-369.

Geary, David M. (2004). The Origin of the Mind: Evolution of Brain, Cognition, and General Intelligence. American Psychological Association (APA).

Gibb, J. R. (1951). The effect of group size and of threat reduction upon creativity in a problem
solving situation. American Psychologist, 6, 324.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, 26, 213–224.

Gottfredson, L.S. (1997). "Foreword to "intelligence and social policy" Intelligence 24 (1): 1–12.

Hackman, J. R., & Morris, C. G., (1975) Group Tasks, Group Interaction Process, and Group Performance Effectiveness: A Review and Proposed Integration. In L. Berkowitz (Ed.) Advances in Experimental Social Psychology, Vol. 8, Academic Press, New York, pp. 47-99,

Hackman, J. R., & Wageman, R. (2005). A theory of team coaching. Academy of Management Review, 30, 269–287.

Hackman, J.R. (2002) Leading Teams: Setting the Stage for Great Performances, Harvard Business Review

Hackman, J.R., & Vidman, N.J. (1970). Effects of size and task type on group performance and member reactions. Sociometry, 33(1), 37-54.

Hare, A.P. (1952). A study of interaction and consensus in different sized groups. American Sociological Review, 17, 261-267

Hare, A.P. (1976) Handbook of Small Group Research, 2nd Ed, New York: Free Press.

Hawkins, C. (1962). Interaction rates of jurors aligned in factors. American Sociological Review, 27, 689-691.

Hertel, G., Geister, S., & Konradt, U. (2005). Managing virtual teams: A review of current empirical research. Human Resource Management Review, 15, 69–95.

Horowitz, I.A., Bordens K.S., (2002) The effects of jury size, evidence complexity, and note taking on jury process and performance in a civil trial. Journal of Applied Psychology, 87 (1),pp. 121–130

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. Experimental Economics, 14, 399–425.

Ingham, Alan G.; Levinger, George; Graves, James; Peckham, Vaughn. (1974) "The Ringelmann effect: Studies of group size and group performance". Journal of Experimental Social Psychology 10 (4): 371–384.

Ipeirotis, P. G (2010, March 9) The New Demographics of Mechanical Turk. Behind the Enemy Lines. Retrieved from http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html

Jackson, J. M. & Harkins, S. G. (1985). "Equity in effort: An explanation of the social loafing effect." Journal of Personality and Social Psychology, 49, 1199-1206

Karau, S.J., Kelly, J.R., (1992). The effects of time scarcity and time abundance on group performance quality and interaction process. Journal of Experimental Social Psychology 28, 542–571.

Karau, Steven J.; Williams, Kipling D. (1993). "Social loafing: A meta-analytic review and theoretical integration". Journal of Personality and Social Psychology 65 (4): 681–706

Kees, J., Christopher B., Scot B., & Sheehan K., (2017), "An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk," Journal of Advertising, 46 (1), 141–55

Kees, J., Scot B., and Heintz Tangari A.(2010), "The Impact of Regulatory Focus, Temporal Orientation, and Fit on Consumer Responses to Health-Related Advertising," Journal of Advertising, 39 (1), 19–34.

Kelley, H., & Thibault, J. (1954). Experimental studies in group problem solving processes. In G. Lindzey (Ed.), Handbook of social psychology (Vol. 2, pp. 735-785). Reading, MA: Addison-Wesley

Kelly, J.R., Karau, S.J., (1999). Group decision making: the effects of initial preferences and time pressure. Personality and Social Psychology Bulletin 25, 1342–1354.

Kelly, J.R., Loving, T.J., (2004). Time pressure and group performance: exploring underlying processes in the attentional focus model. Journal of Experimental Social Psychology 40, 185–198.

Kelly, J.R., McGrath, J.E., (1985). Effects of time limits and task types on task performance and interaction of four-person groups. Journal of Personality and Social Psychology 49, 395–407.

Kerr, N. L., & Bruun, S. E. (1983) The dispensability of member effort and group motivation losses: Free-rider effects. Journal of Personality and Social Psychology, 44, 78.94.

Kosara, R., & Ziemkiewicz, C. (2010, April). Do Mechanical Turks dream of square pie charts? Paper presented at the Proceedings of the 3rd BELIV'10 Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, Atlanta, GA.

Kravitz, David A.; Martin, Barbara (1986). "Ringelmann rediscovered: The original article". Journal of Personality and Social Psychology 50 (5): 936–941. doi:10.1037/0022-3514.50.5.936. ISSN 1939-1315.

Larson, J.R. (2009) In search of synergy in small group performance. Psychology Press

Latane, B., Williams, K., Harkins, S.,Many hands make light the work: The causes and consequences of social loafing. Journal of Personality & Social Psychology, 37 (6) (1979), pp. 822–832

Latané, Bibb; Williams, Kipling; Harkins, Stephen (1979). "Many hands make light the work: The causes and consequences of social loafing". Journal of Personality and Social Psychology 37 (6): 822–832

Laughlin, P. R., (1980) Social Combination Processes of Cooperative, Problem-Solving Groups as Verbal Intellective Tasks. In M. Fishbein (Ed.), Progress in Social Psychology (Vol. 1). Erlbaum, Hillsale, NJ

Legg, S. and M. Hutter. www.vetta.org. A collection of definitions of intelligence, 2006.

Liden, R.C. , Wayne, S.J. , Jaworski, R.A. , Bennett N., Social loafing: A field investigation. Journal of Management, 30 (2) (2004), pp. 285–304

Littlepage, G., and Silbiger, H. (1992). Recognition of expertise in decision-making groups. Small Group Research, 23, 344-355.

Malone, T. W., and Bernstein, M. S. (Eds.) (2015) Handbook of Collective Intelligence. Cambridge, MA: MIT Press, 2015.

Mao A, Mason W, Suri S, Watts DJ (2016) An Experimental Study of Team Size and Performance on a Complex Task. PLoS ONE 11(4): e0153048.

Marvit, M. Z. (2014), "How Crowdworkers Become Ghosts in the Digital Machine," The Nation, February 5, http://www.thenation.com/article/ 178241/how-crowdworkers-became-ghosts-digital-machine

Mason, W., & Suri, S. (2012), "Conducting Behavioral Research on Amazon's Mechanical Turk," Behavioral Research, 44 (1), 1–23.

McCrae R. R., Costa P. T., (1987) Validation of the Five-Factor Model of Personality Across Instruments and Observers. Journal of Personality and Social Psychology 52, 81-90

McGrath, J.E., (1984) Groups: Interaction and Performance. Prentice-Hall, Inc, Englewood Cliffs, N.J.

Menon, T. and Phillips, K., (2008) Getting Even vs. Being the Odd One Out: Conflict and Cohesion in Even and Odd Sized Groups. IACM 21st Annual Conference Paper.

Mueller, J.S. (2012), Why Individuals in Larger Teams Perform Worse. Organizational Behavior and Human Decision Processes, 117 (1), 111-124.

Neisser, U.; Boodoo, G.; Bouchard Jr, T.J.; Boykin, A.W.; Brody, N.; Ceci, S.J.; Halpern, D.F.; Loehlin, J.C.; Perloff, R.; Sternberg, R.J.; Others. (1998) "Intelligence: Knowns and Unknowns". Annual Progress in Child Psychiatry and Child Development.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N., (2009), "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power," Journal of Experimental Social Psychology, 45 (4)

Orasanu, J., & Fischer, U. (1997). "Finding decisions in natural environments: The view from the cockpit" In C. E. Zsambok & G. Klein (Eds.), Naturalistic decision making expertise: Research and applications (pp. 343–357). Mahwah, NJ: Erlbaum.

Paolacci, G., & Chandler, J. (2014). Inside the Turk. Current Directions in Psychological Science, 23(3), 184-188.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5, 411–419.

Pentland A., (2008) Honest Signals: How They Shape Our World. Bradford Books, Cambridge, MA.

Perloff, R.; Sternberg, R.J.; Urbina, S. (1996). "Intelligence: knowns and unknowns". American Psychologist 51.

Peterson, R., & Merunka D. (2014), "Convenience Samples of College Students and Research Reproducibility," Journal of Business Research, 67 (5), 1035–41.

Pontin, J. (2007, March 25). Artificial intelligence: With help from the humans. The New York Times. Retrieved from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

Raven, J. C. (1936). Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. MSc Thesis, University of London.

Riemer, K., & Richter, A. (2010). Tweet inside: Microblogging in a corporate context. Proceedings of the 23rd Bled eConference, 1–17.

Riemer, K., Richter, A., & Bohringer, M. (2010). Enterprise microblogging. Business & Information Systems Engineering, 2(6), 391–394.

Ringelmann, M. (1913) "Recherches sur les moteurs animés: Travail de l'homme" [Research on animate sources of power: The work of man], Annales de l'Institut National Agronomique, 2nd series, vol. 12, pages 1-40.

Ringelmann, M. (1913). Research on animate sources of power: The work of man. Annales de L'Institut National Agronomique, 2e serietome (pp. 1–40).

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers?: Shifting demographics in Mechanical Turk. Paper presented at the CHI'10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, GA.

Schneider, C., & Zimet, C. (1969). Effects of group size on interaction in small groups.

Schuman, H. & Graham K. (1985), "Survey Methods," in Handbook of Social Psychology, Vol. 1, Gardner Lindzey and Elliot Aronson, eds., New York: Random House, 635–97.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. Clinical Psychological Science, 1, 213–220.

Shaw, M. E. (1973) Scaling Group Tasks: A Method for Dimensional Analysis. JSAS Catalog of Selected Documents in Psychology.

Shaw, M. E. (1976). Group dynamics: The psychology of small group behavior (2nd ed.). New York: McGraw-Hill

Sheehan, K. & Pittman M. (2016), The Academic Researcher's Guide to Mechanical Turk, Irvine, CA: Melvin and Leigh.

Slater, P. (1958). Contrasting correlates of group size. Sociometry, 21, 129-139.

Smith, S. M., Roster, C. A., Golden L. L., & Albaum G. S. (2016), "A Multi-Group Analysis of Online Survey Respondent Data Quality: Comparing a Regular USA Consumer Panel to MTurk Samples," Journal of Business Research, 69 (8), 3139–48

Spearman C. (1904) 'General Intelligence', Objectively Determined and Measured. The American Journal of Psychology Vol. 15, No. 2, pp. 201-292

Stasser, G., Titus, W., (1985) Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion Journal of Personality and Social Psychology, 48(6), 1467-1478

Steiner, I. D. (1972). "Group process and productivity". San Diego, CA: Academic Press.

Stewart, G.L., (2006) A meta-analytic review of relationships between team design features and team performance. Journal of Management, 32 (1), pp. 29–55

Stokes J. P., (1983) Components of Group Cohesion. Small Group Research 14, 163.

Straus, S. G. (1999) Testing a Typology of Tasks: An Empirical Validation of McGrath's (1984) Group Task Circumplex. Small Group Research 30, 166-187.

Thompson, L. (2003). Improving the creativity of organizational work groups. The Academy of Management Executive, 17(1), 96–109.

Valacich, J. S., Dennis, A. R., & Connolly, T. (1994). Idea Generation in Computer-Based Groups: A New Ending to an Old Story. Organizational Behavior and Human Decision Processes, 57(3), 448-467

Wageman R., Hackman J. R., Lehman E., (2005) Team Diagnostic Survey: Development of an Instrument. Journal of Applied Behavioral Science 41, 373-398

Wanous, J., & Youtz, M. (1986). Solution diversity and the quality of group decisions. Academy of Management Journal, 29, 149-159.

Wechsler, D. (1944). "The measurement of adult intelligence" (3rd ed.). Baltimore: Williams & Wilkins.

Wicker, A.W., Mehler A., Assimilation of new members in a large and a small church. Journal of Applied Psychology, 55 (2) (1971), pp. 151–156

Wonderlic E. F., Hovland C. I., (1939) The Personnel Test: a restandardized abridgment of the Otis S-A test for business and industrial use. The Journal of Applied Psychology, No. 23.

Woolley A.W. (2009), Means versus Ends: Implications of outcome and process focus for team adaptation and performance. Organization Science 20, 500-515

Woolley, A.W. & Malone, T.W. (2011). What makes a team smarter? Harvard Business Review, June, 32-33.

Woolley, A.W., & Hashmi, N. (2013). Cultivating collective intelligence in online groups. In P. Michelucci (Ed.) Handbook of Human Computation. New York, NY: Springer Science+Business Media.

Woolley, A.W., Aggarwal, I., & Malone, T.W. (2015). Collective intelligence in teams and organizations. In T.W. Malone & M. Bernstein (Eds.) Collective Intelligence. Cambridge, MA: MIT Press.

Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., & T.W. (2010) Evidence for a Collective Intelligence Factor in the Performance of Human Groups. 22222Science 29 October 2010: 330 (6004), 686-688.

Yetton, P. & Bottger, P. (1983) The relationship among group size, member ability, social science decision schemes, and performance. Organizational Behavior and Human Performance, 32(2), 145 - 159