# Are Mock-Jurors Sensitive to the Benefits of Collective Decision-Making?

Emma Tiggemann

*This thesis is submitted in partial fulfilment of the Honours degree of Bachelor of Psychological Science*

School of Psychology

University of Adelaide

September 2021

Word Count: 9356

**Table of Contents**

List of Figures

**Abstract**

Jurors are often presented with forensic expert testimony and tasked with understanding and making decisions about the evidence. Collective decision-making such as blind verification is used due to the benefits of aggregating independent decisions to increase accuracy and reduce potential biases. The current study is the first to investigate how mock jurors interpret the credibility of collective decision-making evidence. We used a 2x4 between-subjects factorial design to test if the number of experts or the independence of their decision affects mock jurors' credibility ratings of the testimony. It was predicted that credibility ratings would increase with group size for both groups, but the independent group would be rated higher in credibility. A student sample and an experienced jury sample were collected to test if this had any effect on the results, then an exploratory meta-analysis was run with the combined sample. Generalised linear models with polynomial contrasts found no significant effects for independence, while only the combined sample showed a significant effect for crowd size on credibility ratings. Results suggest that mock jurors are not sensitive to the benefits of independence in collective decision-making, while they may be sensitive to the number of experts. Further research is required to understand the relationship between these variables and credibility ratings. The use of collective decision-making in forensic science may need to be used with caution if jurors are not as sensitive to the increases and decreases in accuracy that come with collective decision-making as expected.

*Keywords*: Aggregation, verification.

# Declaration

This thesis contains no material which has been accepted for the award of any other degree of diploma in any University, and, to the best of my knowledge, this thesis contains no material previously published except where due reference is made. I give permission for the digital version of this thesis to be made available on the web, via the University of Adelaide's digital thesis repository, the Library Search and through web search engines, unless permission has been granted by the School to restrict access for a period of time.

## Contribution Statement

I worked in collaboration with my supervisor to design, plan and run this study. I conducted the literature search and applied for low-risk ethics through University of Adelaide and received ethics approval prior to commencing data collection. I created the case reports and accompanying visual aids, as well as the survey to be administered, with input from my supervisor. I advertised and collected the first sample of student participants through the SONA research participation system, my supervisor advertised the study on Prolific to collect data for the second sample, pre-screened for prior jury duty. Both I and my supervisor exported the data, checked screened for duplicates and incomplete attempts, formatted it for data analysis, and ran statistical analyses, including generating plots and figures. My supervisor provided the R Markdown file for data analysis, and I wrote all aspects of the thesis.

## Acknowledgements

I would like to acknowledge and thank my supervisor and my family for their support throughout the process of this year.

# Introduction

## 1.1 Forensic Science

In forensic science, the understanding and interpretation of testimony is vital as juries are required to use this information to make decisions about guilt or innocence. Expert witnesses are used to provide jurors with evidence from fields they are unfamiliar with in order to understand and make decisions about disputed facts (Brewer, 1998; Vidmar, 2005). The National Academy of Sciences (NAS) Report (2009) indicated that forensic sciences require more research to decrease errors that come with human decision-making in this field. One way of increasing accuracy in forensic science is the use of collective decision-making or 'verification' of decisions made by examiners. Verification is a part of the ACE-V (analysis, comparison, evaluation, and verification) examination method, an established method for comparing and identifying two items such as fingerprints or facial identification (Vanderkolk, 2011). The ACE-V method was created to parallel the process of the scientific method used for hypothesis testing and objectivity in problem-solving (Reznicek et al., 2010). The ACE procedure, which was first discussed in 1959 by Roy Huber, was then updated by David Ashbaugh in the 1980s with the addition of verification to the process (Speckels, 2011). An integral part of the scientific method and the ACE-V examination method is the peer review or verification process that allows for independent review to test the validity of research or decisions (Scott, 2007). The use of collective decision-making in the verification step such as multiple blind verifications helps to increase the accuracy of forensic science decisions. Despite the use of ACE-V in forensic examinations and the understanding of the importance of independent verification, existing research is limited in understanding how presenting verification information to jurors affects their beliefs regarding the credibility of the evidence. There have been some attempts to improve the presentation of expert testimony to help jurors understand and accurately interpret the quality of the evidence. Unfortunately, presenting statistical explanations of forensic evidence such as likelihood ratios and random match probability have been found to cause more confusion and incorrect interpretation of the evidence (Martire et al., 2013). Other theories about the level at which juries understand and believe experts have also found conflicting results. For example, the CSI effect was a common belief that television shows about crime present forensic evidence as perfect and instantaneous causes laypeople to over-believe experts in real life (Baskin & Sommers, 2010). However, Shelton (2008) found that while CSI viewers had higher expectations of forensic sciences, it did not relate to their conviction decisions. Ribeiro et al., (2019) also noted that the CSI effect was not supported in their study, and in fact, participants were more sceptical of certain forensic techniques than necessary. More research is required to improve forensic science procedures and presentation to understand how evidence that is presented to jurors is understood and used to make decisions. This thesis investigates how two important factors of collective decision-making used in verification, group size, and independence, influence how mock jurors rate the credibility of the evidence presented.

**1.2 Collective Decision-Making Research**

The use of collective decision-making in verification allows for multiple experts to assess the evidence and make their own decision, to increase accuracy and reduce errors (Tindale et al., 2013). In 1906, Francis Galton firmly believed that a group could not be smarter than an individual and conducted an experiment to test his hypothesis. He collected the answers from a competition where people estimated the weight of an ox, then averaged their guesses. He found that the final aggregated result was 1197 pounds, only one pound away from the true weight of the ox (Surowieki, 2005). The almost perfect result of the aggregated guesses was the complete opposite of what Galton expected to see, but it became known as the wisdom of crowds or collective decision-making. Galton was then forced to defend the wisdom of the crowd theory because doubts emerged, similar to his own beliefs originally, that a random group could not be that accurate. One doubt raised was that in order to get that result the competition participants must have been experts, but Galton refuted this stating that the group of approximately 800 people could not have been all experts. Despite not being able to prove how many were experts because this information was not collected, he explained his findings by comparing it to the number of people who bet on horse racing or vote that are not guided by expertise (Galton, 1907). More negative assumptions about group decision-making endured because of the notion that people 'go mad' in groups and act differently than they would individually, for example, that juries would make decisions that no juror would individually approve of (McPhail, 1989). This belief that not only are groups not smarter than individuals but that they go mad comes from the anonymity and loss of responsibility for actions that large numbers provide groups which can lead to dangerous outcomes such as riots (McPhail, 2006; Sorensen, 2019). Research has shown that this is not the case for decision-making groups, and in fact, when used correctly, collective decision-making can increase the accuracy of the decision beyond individual member's abilities. Due to the importance of accuracy in forensic science and the use of blind verification, the current study will address how collective decision-making information is interpreted by mock jurors in a forensic science context.

While the negative beliefs made about group decision-making might have influenced the spread and application of wisdom of the crowd early on, research has established it is beneficial for improving accuracy. Suroweiki (2005) noted that group decisions are best when they meet four criteria that maximise the benefits and improve accuracy, these include diversity, decentralisation, aggregation method, and independence of decisions. Decentralisation and diversity both address the ability of members of the crowd to provide local knowledge and offer different points of view (Hosseini et al., 2015). The independence of decisions before aggregation is important because if people are aware of others' decisions or guesses, it may influence theirs and skew the aggregated result (Langenburg et al., 2009). With the rise in understanding, the application of wisdom of the crowd decision-making can be seen in many areas. Wisdom of the crowds is commonly used in prediction markets, showing the ability of crowds to outperform polls or expert opinions in elections, sports betting or stock markets (Matzler et al., 2016; Murr, 2011). Further research has found that the application of wisdom of the crowd decision-making is being used for small- and large-scale enterprises as it can be applied to a range of groups and problems (Hosseini et al., 2015). Due to the evolution of what a crowd can entail, wisdom of the crowd decision-making is even the basis of juries as it uses the same concepts of a diverse group of laypeople seeking the truth (Sorensen, 2019). The method used for aggregation has also evolved with the uses of wisdom of the crowd, from simple averaging to majority rules for when decisions are not estimates, while still improving accuracy (Sorkin et al., 1998). Through the expansion of understanding how wisdom of the crowd decision-making can be used, research has found that it can be extended beyond groups of laypeople, to groups of experts to increase accuracy in critical areas. For example, wisdom of the crowd has been used in the medical field with groups of clinicians and radiologists (Kattan et al., 2016; Wolf et al., 2015). Wisdom of the crowd decision-making has also shown improvements in accuracy for forensic sciences such as face matching and fingerprint identification (Tangen et al., 2020; White et al., 2013).

In their study on the use of wisdom of the crowd for experts in fingerprint matching, Tangen et al., (2020), found that the improvements in accuracy was obvious for both expert and novice groups and increased with group size. This improvement was, however, better for the expert groups and when they had unlimited time to make the decisions, pooling their decisions decreased false positive rates by up to 8% and false-negative rates by up to 12%. They also tested rules for the most accurate aggregate decision, which included taking the majority, following the most confident group member, or the most senior member. They found that the majority rule provided the largest increase in accuracy for both the expert and novice groups. This is consistent with previous research into the best use of majority rules for group decision-making. Sorkin et al., (1998) compared simple majority rules to more stringent rules such as taking the three-quarters majority or unanimous voting, finding that the normal majority rule is the best option as more liberal criteria was used by group members for more stringent majority rules, which would not be ideal for use in forensic science. White et al., (2013) used wisdom of the crowd effects in unfamiliar face matching and also found that groups of four and above showed increases in accuracy, also using the majority rule. Even when looking at groups of low-performing individuals, the aggregated group of eight was better than the most accurate individual, indicating that the increases in accuracy are not only due to high-performing individuals in the groups. White et al., (2013) also tested another technique for aggregating group decisions by collecting ratings of similarity for the faces rather than a binary same or not same decision. This also showed increases in accuracy with group size, to near perfect at groups of eight, then hitting a ceiling effect as it levelled out. Based on this research the current study addresses the increases in accuracy up to groups of nine experts to see if mock-jurors are sensitive to this through credibility ratings.

A critical part of verification and the wisdom of the crowd effect is the independence of individuals' decisions before aggregation (Vanderkolk, 2009). This is referred to as blind verification which is used to avoid any contextual bias that may affect the expert's interpretation of the evidence. These can include confirmation bias, pressure to agree with the initial expert, information about the case, reference materials for identifications, and more (Dror, 2015). Langenburg et al., (2009) noted that without the use of independence in testing, if subsequent scientists are making decisions based on the original expert's decision, they may be more likely to miss or discount obvious errors to verify rather than disconfirm the decision. It is important to allow verifying experts the ability to make their own individual, and potentially different identifications to avoid biases and influences, this is hard at times when they are seen as negative rather than informative results (Dror, 2015). Social influence can completely undermine the benefits of wisdom of the crowd decision-making by reducing the diversity of the crowd without reducing the errors and increasing the confidence of decision-makers (Lorenz et al., 2011). Overconfidence in experts is problematic because confidence in forensic expert testimony can influence the credibility that the jury attributes to them and the evidence presented (Cramer et al., 2009). Not only is this significant for the accuracy of identifications but if jurors are not aware of these potential issues with non-independent verification, they may interpret it as more accurate than it is.

Independence is important in the process of forensic science and identifications, and it has also been found to be important to judges, lawyers, and jurors. According to a study by Shuman et al., (1996), expert credibility was influenced by the expert's impartiality or the perceived impartiality by the juror, indicating that they are sensitive to the importance of independence. The American Bar Association (ABA) created a committee to assess jury comprehension of complex cases. They also found that independent involvement made the expert witnesses more believable. Similarly, when experts did their own independent research and were not just 'hired guns' to agree with one side, they were viewed as more impartial and believable (ABA committee, 1989, as cited in Champagne et al., 1991). Shuman et al., (1996) also noted that along with experts being familiar with the case and having clear communication of their testimony, independence was one of the largest influences on how believable the jurors thought they were (Shuman et al., 1996). These studies on real judges, lawyers, and jurors show that there is an understanding that independence is important in the forensic science context, but it is not known if jurors are sensitive to this in group decision-making such as blind verification.

Increasing accuracy in forensic science is necessary because of the potentially devastating outcomes of wrongful convictions or false acquittals. The innocence project and similar studies around the world have acquitted an alarming number of people through DNA evidence and flawed expert testimony contributed to a significant number of these false imprisonments (Innocence Project, n.d.). Given the influence that expert testimony has it is even more urgent to understand how laypeople interpret and use this information when making their own decisions. Despite some beliefs, human decision-making is deeply embedded into all facets of the justice system including forensic science, eyewitness memory, and judge and jury decision-making (Dror, 2015). Even with highly accurate forensic science disciplines such as fingerprint matching where computers can help provide a list of potential matches, ultimately a decision has to be made and this human aspect includes some level of subjectivity and potential for inaccuracy (Behson & Koppl, 2013). Research has been conducted to test expertise in forensic scientists and while fingerprint experts were found to be significantly more accurate than novices, they are not infallible (Tangen et al., 2011). This human factor needs to be addressed for forensic science to be used as optimally as possible (Dror, 2015). Jury members are required to interpret, understand and appropriately weigh evidence from a variety of sources during a trial. Forensic science evidence that has been through the ACE-V examination method is verified by one or more experts and jurors should have all the information to truly understand and use that evidence in their decisions. Therefore, it is important to understand if jurors are sensitive to not only increases but potential decreases in accuracy from the processes of verification, such as non-independent decisions.

The existing area of research most analogous to the current study is how the level of consensus, majority versus minority, influences people's beliefs about the validity of a message. In these studies, majority consensus was associated with higher message validity compared to minority consensus (Bohner et al., 2008; Bohner et al., 1996). This high level of consensus indicates to recipients that it is correct or more likely to be reflective of the objective truth (Kelley, 1967). Darke et al., (1998) found that attitudes of participants who were highly motivated regarding accuracy were more influenced by majority rules with reliably large groups. This suggests that laypeople may indeed be sensitive to consensus and increases of accuracy, specifically when they are motivated, which would be expected of a jury. The current study uses total consensus to test how crowd size and independence of wisdom of the crowd decision-making is interpreted by laypeople in the forensic science context.

**1.3 Credibility Research**

Credibility and persuasiveness are both common measures used in cognitive forensic research to understand how jurors and mock jurors evaluate expert testimony. They are also linked as credibility influences persuasiveness which also impacts jury decision-making outcomes in courtroom proceedings. (Brodsky, 2010; McCarthy Wilcox & NicDaeid, 2018).

Research into what affects credibility ratings of forensic experts has found mixed results due to the myriad of variables that can influence it. Specifically, a large amount of research has focussed on how the attributes of the expert are used to gauge their credibility and therefore the credibility of the evidence they are presenting. For example, expert witness confidence has been found to increase the credibility ratings of experts (Cramer et al., 2009). The Witness Credibility Scale, created by Brodsky et al., (2010) hypothesized that credibility ratings were the product of factors relating to the expert such as confidence, likeability, trustworthiness, and intelligence, which has been supported by other studies. Brodsky et al., (2009) found that ratings of trustworthiness were related to the likeability of the expert. In contrast, Mcarthy Wilcox and NicDaeid (2018) surveyed real jurors to investigate what was considered important in forensic experts. Education experience was noted as the most important compared to laboratory accreditation. Blackwell and Seymour (2015) also found that experience was associated with high credibility compared to just qualifications.

More recently, research has been directed at the effect that the quality of the testimony itself has on credibility ratings. Parrott et al., (2015) varied the level of knowledge of the expert in a mock trial study to see if highly knowledgeable experts are rated more credible, as would be expected. Surprisingly, only likeability increased for the knowledgeable expert but there was no effect on credibility ratings. In contrast, another study found that when scientific data was provided by the expert witness, there was a significant increase in credibility ratings ("The Joint Effect", 2019). Salerno et al., (2017) also tested the influence of strong versus weak expert testimonies on mock juror's individual and group decision-making processes. They found that individually when motivated to focus on substantive information in the testimony, they accurately evaluated the strength of the testimony. Similarly, Martire et al., (2020) studied the influence of strong versus weak expert reports on the persuasiveness of the testimony to test the influence of the Expert Persuasion Expectancy (ExPEx) Framework. The ExPEx framework consists of attributes that influence the persuasiveness of an expert to some extent, including foundation, field, specialty, ability, opinion, support, consistency, and trustworthiness. Martire et al., (2020) found that the strong testimony had a significant effect on persuasiveness in their study, but ability, consistency, and trustworthiness had the most influence on expert persuasion. They suggested that these were logically appropriate because they focus on the proficiency of the expert, endorsements by other experts, and their objectivity. Verification and independence are covered by these highly influential attributes, however, it is not yet known if mock jurors are actually sensitive to manipulation of these variables. Although Martire et al., (2020) used persuasiveness to measure the influence of the ExPEx variables, they also noted that credibility is an outcome variable that is appropriate for capturing the quality of expert testimony rather than a contributing factor. This thesis will expand on the influence expert testimony has on credibility ratings to understand if mock jurors are sensitive to the increases in accuracy that come with wisdom of the crowd decision-making and the independence of decisions.

As mentioned above, there are numerous variables known to influence the credibility and persuasiveness ratings of expert witnesses, so studies have used a range of measures. The witness credibility scale (Brodsky et al., 2010) is a measure of source credibility that focusses on the specific characteristics of the expert to measure credibility. Martire uses a measure of persuasiveness instead which includes ratings of credibility, value, and weight of the testimony as rated by participants (Martire et al., 2020; Martire & Montgomery-Farrer, 2020). The current study uses this format to measure the credibility of the testimony rather than the expert, through ratings of credibility, trustworthiness, and weight.

## 1.4 Current Project

Research has demonstrated that wisdom of the crowd decision-making increases accuracy, and some studies have also shown that jurors are sensitive to increases in scientific accuracy ("The Joint Effect", 2019). This could indicate that mock jurors will also be sensitive to the wisdom of the crowd effects. Based on these findings and the belief that consensus implies correctness (Bohner et al., 2008), we are expecting to see a main effect of the first independent variable, the number of experts, on credibility ratings. So far, research on the use of independence in wisdom of the crowds for forensic science and how it affects credibility is limited. There have been mixed results in other areas due to the complexity of group decision-making across disciplines. Some research has shown that non-independence or interaction between crowd members could actually improve accuracy in some problem-solving cases (Jeckeln et al., 2018; Navajas et al., 2018). Whereas in a lot of studies on the wisdom of crowds, independence is a requirement or an assumption of the crowd (Kurvers et al., 2021; Sorkin et al. 1998). Despite the varying use of independence in these areas, in forensic science, independence and impartiality have been found to make experts more believable (ABA committee, 1989, as cited in Shuman et al., 1996). Champagne et al (1991) also found that independence was one of the key impacts on the quality of the expert testimony as rated by jurors. While it has not yet been applied to the forensic science context, lack of independence in wisdom of the crowd decision-making is known to affect the accuracy of the group due to social influence (Lorenz et al., 2011). Additionally, jurors do seem to hold the belief that independence is good as it is associated with reliability and impartiality. Due to this, we are expecting to see a main effect of independence on the credibility ratings of the evidence.

The current project investigates whether and to what extent mock jury members are sensitive to the benefits of multiple independent expert judgments when weighing up the credibility of forensic science evidence. While verification is already used to some extent in forensic contexts, it is unknown how presenting this information actually influences the jury. This study explores the effect of increasing the number of verifying experts on mock juror's ratings of credibility. The effect of independence or lack of independence between experts' decisions on credibility ratings is also investigated due to the importance of this factor on the wisdom of the crowd effect. Participants are presented with four forensic experts who have made an identification which is then verified by 0, 2, 4, or 8 other experts independently or while knowing the original expert's decision. Mock juror participants rate the credibility, weight, and trustworthiness of the evidence presented to them in each case report for an aggregated measure of credibility. We hypothesize that credibility ratings increase with the number of experts for both levels of independence, but the non-independent group will be rated lower overall in credibility.

There is a range of research and criticism about the generalisability of mock-jury research as well potential influences of prior jury experience. Studies have shown mixed results about whether student mock jurors interpret information the same way as jury-eligible mock juror samples. Some indicate that any differences found are negligible and that they respond similarly to court proceedings to real jurors (Bornstein, 1999; Bornstein et al., 2017; MacCoun 1989). In contrast, more recent research has found differences for student and community mock juries in judgements of culpability, persuasion by expert testimony, and cognitive processing style which is related to verdicts reached by jurors (Keller et al., 2011; McCabe et al., 2010). The effects of prior jury experience have been debated in the literature as well as it has been found to influence sentencing (Himelein et al., 1991) but not on verdicts (Werner et al., 1985). Due to this assortment of results, we decided to replicate our study with two participant pools to compare the effect that prior jury experience and a more representative group might have on the results. Experiment 1a is a sample of first-year University of Adelaide psychology students, while Experiment 1b is an adult sample of participants pre-screened for prior experience participating in jury duty.

# Experiment 1a

This study used a 4(Crowd Size: 1,3,5,9) x 2(Independence: Independent, Dependent) fully between-subjects factorial experimental design to examine the effect of the number of forensic expert verifications as well as the independence of their decisions on mock juror credibility ratings. The number of experts, 1,3,5, or 9, were decided upon based on the study aggregating fingerprint expert's decisions by Tangen et al., (2020). This study showed large increases in accuracy up until five experts, with smaller increases to seven and nine experts, which was also approaching the asymptote for increases in accuracy. A study on face identification experts also found that the improvement in accuracy was saturated by a group size of eight (Balsdon et al., 2018). This showed that it was not worth using group sizes above nine as large increases in accuracy do not continue and would therefore not be useful in forensic science.

## 2.1 Method

**2.1.1 Participants.** I advertised the study through the online recruitment platform SONA for first-year psychology students from the University of Adelaide and granted each participant course credit through the system as compensation for their time. Participants met the requirements of English proficiency, but students aged under 18 were also accepted. This study was approved by the Research Ethics Committee at the University of Adelaide, ethics approval number: 21/42. Participants were provided with an information sheet (Appendix 1) that outlined any potential risks, the ethics approval number, and contact details for any questions regarding the study or research, as well as a consent form at the start of the online survey (Appendix 2). The information sheet also noted that the data will be de-identified to ensure anonymity and they could withdraw any time before the survey is submitted.

In order to estimate an appropriate sample size a priori, I conducted a power analysis. A range of effect size estimates from studies on jury assessment of forensic experts were considered. Effect sizes found in the studies considered for the power analysis ranged from Cohen's d of .014 to 1.01(Koehler et al., 2003; Martire et al., 2020; Parrott et al., 2015). A study by Martire and Montgomery-Farrer (2020) also found significantly large effect sizes above 2 on the effect of high- versus low-quality expert testimonies. For the current study, a moderate effect size estimate of .5 was used for the power analysis as it falls within this range of reported effect sizes while erring on the side of minimising misses. The power analysis indicated that a minimum of 18 participants per group (a total of 144 participants) would be required to detect an effect size of .5 with approximately 84% power.

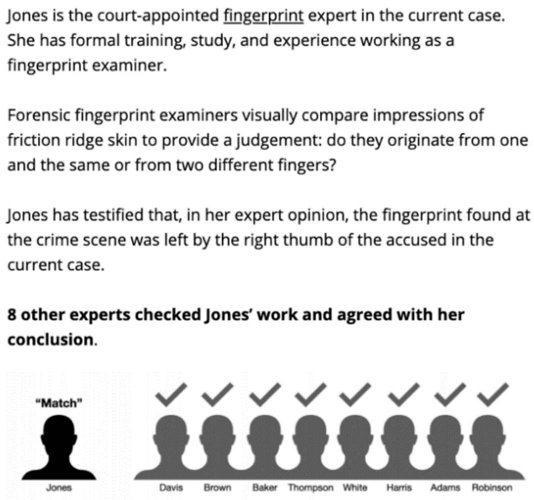
By the data collection cut-off date, the 17$^{th}$ of August 2021, a total of 138 participants had completed the study. Ultimately, group sizes varied from 16 to 18 participants. Participants were aged between 17 and 43, (median = 18, mean = 19.8). The sample consisted of 24 males, 113 females and one preferred not to say. A majority of participants (94) had completed high school, 33 had some university experience but no degree, four had less than a high school degree, three had a completed a bachelor's degree already, one had a master's degree, two an associate degree, while one preferred not to say. Only three participants stated that they had previously done jury duty.

**2.1.2 Case Reports.** To increase the generalisability and power of the study, the manipulation was presented with four expert reports from different forensic science disciplines. This was also to ensure credibility ratings were based on the manipulation rather than just highly accepted and believed forensic sciences. The forensic sciences used in these reports were fingerprint examination, hand-writing analysis, dental examination, and facial image analysis.

These were selected as each can be used to make identifications of a suspect to allow the format to be the same between reports. The four case reports were kept simple and formatted the same for each report, except for the forensic discipline and type of evidence described, they were also the same across the eight conditions except for the manipulation. The reports were constructed to include a brief introduction of the main forensic expert making the testimony, an explanation of the forensic science of their expertise, and the conclusion made by the expert. The manipulation was presented in a written statement as well as visually to ensure it was clear, examples of the case reports are shown in Figure 1. Research has shown that visualisations help jurors comprehend information (Wilcox & NicDaeid, 2018), and written statements of testimonies also help with comprehension (ForsterLee et al., 2000). Visualisations were also used to protect from the potential limitation of participants not fully reading the reports and make the manipulation salient. The names of the main experts and the verifying experts were also presented to make it obvious that each are individual experts, whether their decision was independent or not.

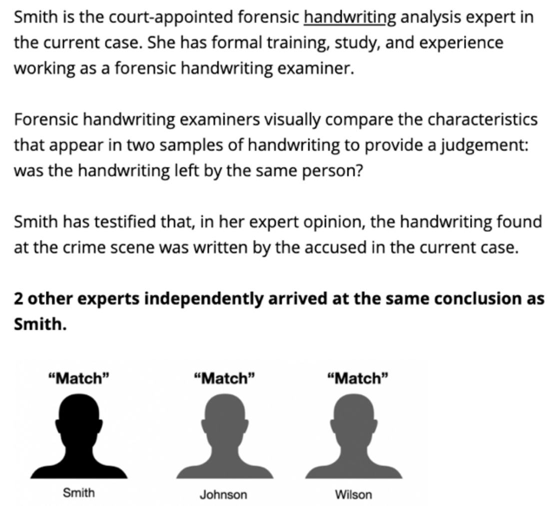
**Figure 1**

(a) *Examples of case reports*




*Note*. This is an example of the fingerprint report presented for the non-independent group with 9 experts, the 8 'verifying' experts are grouped together to show they are confirming the original expert's decision, the manipulation is stated in the bolded line.

(b)




*Note*. This is an example of the handwriting analysis report presented for the independent group with 3 experts, the 2 verifying experts are separate to show their independence, the manipulation is stated on the bolded line.

**2.1.3 Measures.** Credibility assessment is a common measure in cognitive forensic studies with real as well as mock jurors (Brodsky et al., 2010; Cramer et al., 2009). To measure credibility, participants were asked to rate the evidence presented to them. Participants rated the credibility of the evidence, its trustworthiness, and the weight they would put on the evidence on a scale from 0 (not at all) to 100 (completely). The scale of 100 was used to ensure a large range of credibility values could be captured, including 0 if they believe there is no credibility. These ratings are then aggregated for a total measure of credibility due to potentially ambiguous definitions or understandings of credibility. Studies have shown that weight and credibility are highly positively correlated (Choo, 1964; Martire et al., 2020), and trustworthiness is a key component of expert witness credibility (Brodsky et al., 2010). Our measure was based on previous studies that use measures of the persuasiveness of forensic experts by aggregating credibility, value, and weight (Martire et al., 2020). Although research has shown that jurors often use personal attributes of experts to assess credibility, the current study focuses on ratings of the credibility of the testimony due to the importance of accuracy and truth in the forensic science discipline (Smith et al., 2020).

Our measure of credibility included ratings of credibility, weight, and trustworthiness. To ensure this was an appropriate measure of credibility, after results were collected the three items were tested for correlations and internal consistency. All three items were significant and highly positively correlated with each other, for trust and weight; $r(136) = .89$, $p<.01$, trust and credibility; $r(136) = .94$, $p<.01$, and weight and credibility; $r(136) = .82$, $p<.01$. Internal reliability was also significantly high (Cronbach's alpha = 0.96).

**2.1.4 Procedure.** The survey was administered online via the Qualtrics platform, which participants completed on their personal computer or device. Participants were randomly assigned to one of the 8 conditions of the survey, where they gave informed consent to begin. They then read the 4 forensic expert case reports and rated the credibility, trustworthiness, and weight they would assign to the evidence for each case report. Participants then answered demographic questions including age, gender, and level of education. Additionally, participants were also asked to state whether they had previous experience doing jury duty.

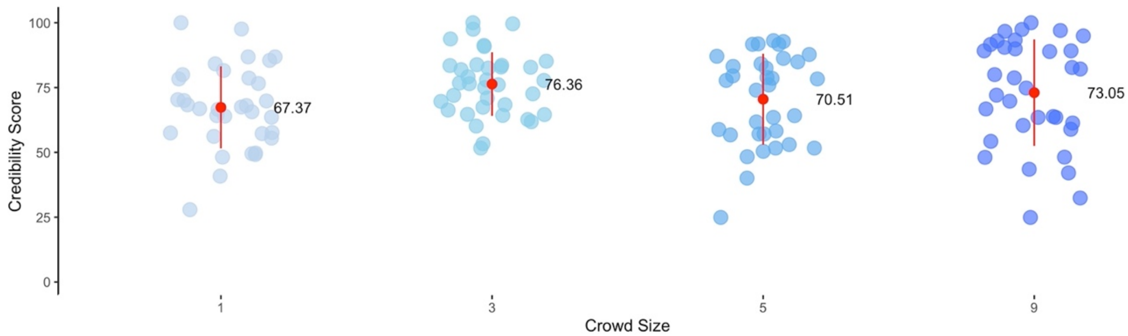
## 2.2 Results

Participants' average completion time of the survey was 4.7 minutes. Each of the reports in the survey did not allow participants to continue until at least 10 seconds had passed, however, they actually spent an average of 36 seconds on these pages, implying that they were paying attention to the report and the manipulation. The data was screened to ensure the criteria were met and to remove incomplete or repeated attempts. Ratings of credibility, trustworthiness, and weight were aggregated for a total credibility rating across all four forensic expert reports. As noted above this was possible due to the high correlations and internal consistency between the items. Graphs of the data were formulated to first examine the distribution and shape of the data in relation to the independent and dependent variables. Figure 2 shows the credibility ratings by crowd size. Planned analyses included a 2x4 between-subjects factorial ANOVA and planned polynomial contrasts.

I conducted a Shapiro-Wilk test, which indicated that the assumption of normality for ANOVA was not met ($p=.016$) As this assumption was not met, the planned 2x4 between-subjects factorial ANOVA was no longer appropriate to test the effect of our variables on credibility ratings. Instead, a generalised linear model with a Gamma distribution was used with a good fit (AIC = 1195.4). The generalised linear model with polynomial contrasts revealed no significant effect of group size on credibility ratings ($p>.05$). No effect of independence was found on credibility ratings either ($p>.05$). The mean credibility scores between the groups show a minimal difference, the independent group was 73.29 while the non-independent group was 70.29. This difference in the credibility ratings of three points was not large enough for a significant effect to be found.

**Figure 2**

*Scatterplot Experiment 1a*




*Note*. Each data point represents individual participants' aggregated credibility score (out of 100) by crowd size (1,3,5,9), showing the mean for each group and the standard deviation in red.

## Experiment 1b

One explanation for the lack of an effect in Experiment 1a is the sample of student mock jurors who are mostly around 18 years old and never completed jury duty before, apart from three. Studies have shown that there are some differences between student mock juries and community juries with jury experience. As McCabe et al., (2010) found that students were not as persuaded by clinical expert testimony as community samples of mock jurors, it is worth testing if the sensitivity to wisdom of the crowd decision information differs between these groups. The next experiment tests this by using a sample of mock jurors pre-screened for prior jury duty, to test for the effects of independence and group size on credibility ratings when multiple experts verify evidence.

### 3.1 Method

**3.1.1 Participants.** My supervisor posted the study on the online recruitment platform Prolific Academic to a second sample of participants who were pre-screened for English proficiency and prior jury duty. A small monetary compensation was provided through Prolific for their time. A total of 145 participants were recruited between the ages of 18 and 65 (Mean = 34.78, Median = 32). Participants consisted of 88 males, 56 females, and one non-binary. This sample was overall quite highly educated, 45 participants had completed a bachelor's degree, 35 had completed a master's degree, six had completed a doctorate, 11 an honours degree, four an associate degree, 21 were high school graduates, 18 had been to college but had not completed a degree, two had a professional degree, two participants had less than a high school degree, and one preferred not to say. Although participants were pre-screened for having previously participated in jury duty, this was not shown by the answers to this question in the survey. 68 participants indicated that they had previously done jury duty, and 77 said they had not, meaning that the pre-screening criteria in prolific may be slightly different from what we used in the survey, however, this sample can still be compared to the student sample as it is a more diverse and experienced, jury-eligible sample.

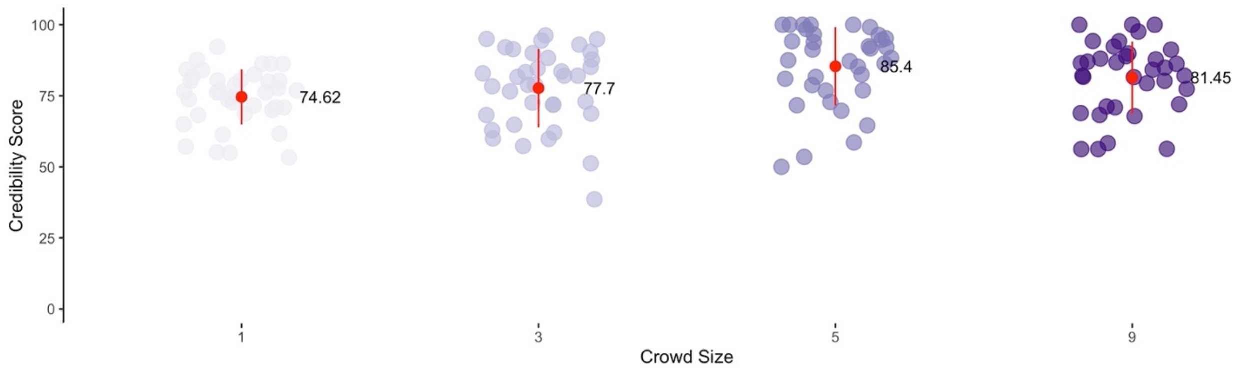
**3.1.2 Measures and Procedure.** The measures, procedure, and experimental design were all replicated exactly from Experiment 1a, with this sample. The median completion time for the survey was 4.3 seconds, but the median time participants spent reading the reports was 45.9 seconds. The three items of the credibility measure were checked to test their correlation and internal consistency. All three items were highly correlated with each other, trust and weight $r(143) = .89$, $p<.01$, trust and credibility, $r(143) = .92$, $p<.01$, and weight and credibility, $r(143) = .85$, $p<.01$. Internal reliability was also significantly high (Cronbach's alpha = 0.96).

### 3.2 Results

Data analysis was run using the same process as Experiment 1a, as it is a replication. Credibility scores based on crowd size were plotted, Figure 3 shows the credibility ratings for each crowd size. The Shapiro-Wilk test was used to test the normality assumption for ANOVA, however, like the first sample, it did not meet the assumption of normality ($p=.000028$). As indicated by the plot in Figure 3 the data was negatively skewed. Therefore, a generalised linear model with a Gamma distribution was used due to the fit (AIC = 1170.7) as well as polynomial contrasts. Surprisingly, this also found no significant effect for crowd size or independence on credibility ratings ($p>.05$). This is clear for independence as the overall mean credibility score for both groups was only 0.2 difference as the independent group was rated 79.5, while the non-independent group was 79.3.

**Figure 3**

*Scatterplot Experiment 1b*




*Note*. Each data point represents individual participants' aggregated credibility score (out of 100) by crowd size (1,3,5,9), showing the mean for each group and the standard deviation in red.

## Meta-Analysis

To address some potential explanations for the lack of significant effects found in Experiments 1a and 1b, we collated the data from the two samples to conduct an exploratory meta-analysis to test for any significant effects of independence or crowd size. Despite running an a priori power analysis to determine the appropriate sample size, the lack of significant results could be due to the study being underpowered due to overestimating the effect size and therefore not having a sufficient sample to detect an effect. Alternatively, it could simply be that the manipulation of the independent variables was not strong enough. To test this, we decided to combine the data from our samples and run an exploratory meta-analysis. Combining the samples also allows for the sample to include a range of experiences regarding previous jury duty, as would be expected in a jury.

### 4.1 Method

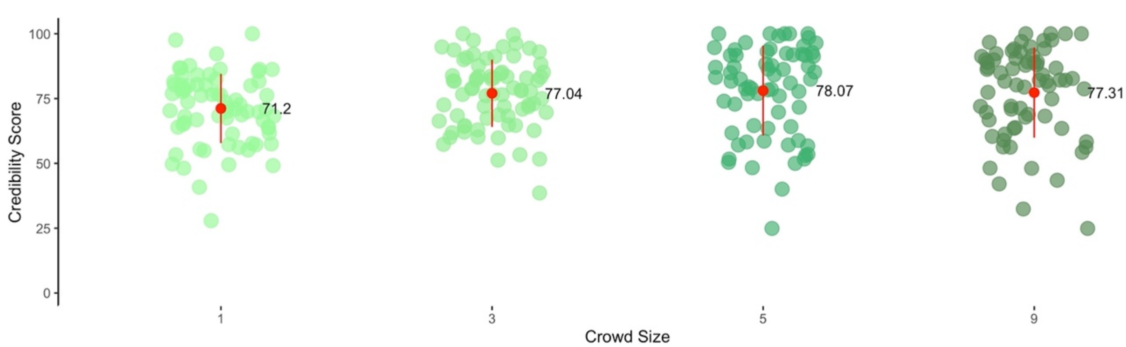
**4.1.1 Participants.** With the samples combined there is a total of 283 participants, consisting of 169 females, 112 males one non-binary and one participant who preferred not to say, the age range is 17 to 65 (median=22). The education reported for this combined sample included 115 who had graduated high school, 51 had some university but not completed it, 48 had completed an undergraduate degree, 11 an honours degree, 36 a master's degree, six had completed an associate two-year degree, six a doctorate degree, six had not completed high school, two had done a professional degree, while two preferred not to say.

### 4.2 Results

The median completion time for the survey was 4.5 minutes, the median time participants spent reading the reports was 41.6 seconds. The three items of the credibility measure were checked to test their correlation and internal consistency. All three items were highly correlated with each other, trust and weight r(281) = .89, p<.01, trust and credibility, r(281) = .94, p<.01, and weight and credibility, r(281) = .83, p<.01. Internal reliability was also significantly high (Cronbach's alpha = 0.96).

Data analysis was run using the same process as Experiments 1a and 1b. Credibility scores based on crowd size were plotted, as seen in Figure 4. The residuals for this data were also tested for normality using the Shapiro-Wilk test but did not meet the assumption of normality (p=.0000012). A generalised linear model with a Gamma distribution was used for this combined sample (AIC = 2399) with polynomial contrasts. This revealed a significant quadratic effect of crowd size on credibility ratings (p<.05). Figure 4 shows that the increase in credibility ratings from the single expert (M=71.2) continues to the three expert group (M=77.04) and five experts (M=78.07) before decreasing slightly for the group of nine experts (M=77.31). No significant effect was found for independence on credibility ratings (p>.05). This is clear as the overall mean credibility score for both groups was 1.74 difference as the independent group was rated 76.76, while the non-independent group was 75.02.

**Figure 4**
*Scatterplot Meta-analysis*




*Note*. Each data point represents individual participants' aggregated credibility score (out of 100) by crowd size (1,3,5,9), showing the mean for each group and the standard deviation in red.

# Discussion

## 5.1 General Discussion

The current research project addresses the gap in cognitive forensic literature regarding how mock jurors interpret wisdom of the crowd decision information in terms of credibility ratings. Two samples completed the same survey and rated the credibility of evidence that either 1, 3, 5, or 9 forensic experts had verified, and the independence of their decisions. Participants rated the credibility of each report, which was then aggregated for a total credibility score to test if the effect of the independent variables was significant. Interestingly, no significant effects were found for independence or crowd size in either sample. This may be informative as it shows that the use of decision information may not necessarily be taken into account when determining credibility in practice. The benefits of blind verification on accuracy are still important but it is also troublesome if mock jurors are not sensitive to this. To test this further, data analysis was run for the two samples combined, showing a significant quadratic effect for crowd size on credibility ratings. However, there was no significant effect for independence for the combined sample either. Overall, our hypothesis was not completely supported by the results of the current study as independent decisions were not rated higher than non-independent decisions and only the combined sample showed a significant effect for crowd size, which was quadratic.

The lack of significant effects found for crowd size on credibility ratings in the two independent samples could be attributed to a number of factors. Results from the current study support the findings from Parrott et al., (2015) that the knowledge of the expert witness does not significantly affect their ratings of the credibility of the evidence. Our study shows that mock jurors may not be significantly sensitive to the increases in accuracy that come with aggregated or verified forensic expert decisions. The significant quadratic effect found when the samples were combined, there is some sensitivity to the benefits of collective decision-making which was not found in Experiments 1a and 1b with smaller samples. Although results from the meta-analysis show a significant quadratic effect, this does not continue to increase with crowd size as expected, it slightly decreases at the largest group of nine experts, which was not predicted. Research does show that increases in accuracy do generally peak around groups consisting of eight or nine experts (Balsdon et al., 2018; Tangen et al., 2020). It is interesting that the decrease in credibility ratings seem to line up with this in Experiment 1b, however it is unlikely that this is due to mock jurors being aware of this ceiling effect for accuracy. An alternative explanation is that the number of experts verifying the evidence presented to participants was met with some scepticism. Participants may have felt that it is unreasonable to believe that nine experts would be assessing the same piece of evidence. Therefore, it may not be worth wasting resources to that extent if more accurate expert testimonies verified by nine experts are rated as less credible.

Although there was no difference between the samples in the main analyses, some differences can be seen in the responses, indicating there might also be some more scepticism in one of the samples. Results show that for the two identical studies with different samples, the variance of credibility ratings is less for the student sample. The experienced jury sample had only 1 participant rate the credibility below 50 (38.58333), whereas the student sample had 15 participants rate credibility lower than 50, the lowest score being 24.91667. As shown in Figures 2 and 3, the mean credibility scores from the student sample are lower, from 67 to 76 within the four groups. In comparison, for the more diverse sample, mean credibility scores for the four group sizes were between 74 and 85. This supports the idea that students may be more sceptical than experienced jurors. Higher credibility scores from the sample that had previously completed jury duty, might be due to them believing more in the credibility of experts due to their experience on jury duty and how the process works. McCabe et al., (2010) compared a student and a community sample on a range of variables and found that the representative sample was persuaded more by clinical expert testimony. Our study focussed on forensic expert testimony but supported these results that the student sample is not as persuaded by experts as a more representative jury sample. This could also be due to the rate at which psychology students participate in psychological research. Learning about research procedures and participating in multiple studies may cause them to be more aware of manipulations and sceptical of them, in this case, not believing that experts are that credible, or that so many experts would make verifications in a real case.

Independence was found to have no significant effect on credibility ratings across the samples, which could have several explanations. While research has shown that people value independence in forensic settings (Champagne et al., 1991; Shuman et al., 1996), mock jurors may simply not be aware of the importance it holds in reducing errors and biases. Conversely, even if mock jurors know that independence as a concept is good, the importance of the application of this to wisdom of the crowd decision-making, or forensic science, may not be obvious to them. For example, research by Cooper & Hall (2000) found that even when mock jurors rated an expert less biased due to being court-appointed, they did not rate them more credible. The current study did not measure belief regarding bias or accuracy of the evidence; however, this could be one explanation for why no effect of independence was found. Therefore, credibility ratings may not provide all information needed to understand how mock jurors comprehend information such as mock juror's sensitivity to certain presentations of evidence. The lack of an effect found for independence also has important implications for some of the beliefs and assumptions in court proceedings regarding the impartiality of testimonies. Studies show that judges, lawyers, and jurors believe that impartiality makes an expert more believable (Champagne et al., 1991), but current results show that it may not necessarily influence their perceived credibility and therefore verdict as would be generally expected.

Limitations of the current study outlined below may have influenced the results as no effect was found for either independent variable on credibility ratings in the individual samples. When data from the two samples were pooled together, there was a significant quadratic effect for crowd size, but independence was still not found to be significant. This might indicate that the lack of an effect of independence is robust and should be researched further to understand why and addressed in practice. In contrast, the fact that there was an effect for crowd size in the pooled sample, could just be because of the increase in sample size, or it may mean that mock jurors are somewhat sensitive to wisdom of the crowd group sizes in experts, however, this would also need to be investigated further to be determined.

**5.2 Limitations**

The main limitation for the interpretation of results from the current study is the collection of data for crowd sizes of only 1,3,5 and 9 experts. This decision was made based on the trend in increasing accuracy found by (Tangen et al., 2020). They showed that the increase in accuracy that comes with collective decision-making for fingerprint identification experts had large increases until five, and then smaller increases after that, then approaching the asymptote after nine experts. The sample collected for Experiment 1a was limited by student participation and the data collection cut-off date. The decision was made to collect the group of nine experts rather than the group of seven to ensure the required sample size required based on the power analysis could be collected while also measuring how participants rated the credibility of small and large groups. Ultimately, this was informative as the group of nine experts decreased in credibility ratings when the samples were pooled. However, only collecting data for those four groups does limit the interpretation of the results and the potential implications of this study, because, without the group of seven experts, the increases in crowd size were not linear. This means that the decrease of credibility ratings between the groups of five and nine experts may have occurred at any point. It is not known based on the data we collected whether the diminishing returns would have started earlier or continued to increase before group nine. Additionally, this may have affected the ability of the results to fit the model.

The sample size for the current study was calculated using an effect size estimate based on research on the expert witness characteristics and testimony that influence ratings of credibility and persuasiveness. As there was a wide range of effects sizes in this research, we used a moderate effect size to estimate the sample size required. Based on the results, a limitation of the current study could be that the effect size was overestimated, underpowering the study, meaning that the sample size may not have been sufficient to measure a significant effect. The student sample did not reach the desired sample size based on the power analysis, while the experienced sample did, however, the results were the same as no significant effects were found. An insufficient sample size could explain the non-significant result found in both the individual samples for both independent variables, while the combined sample showed a significant effect for crowd size on credibility ratings.

Mock juror studies often suffer from limitations regarding the ecological validity of using mock jurors and the format of presenting the information (Bornstein et al., 2017). This is a limitation of the current study as participants received only the relevant case information regarding the expert and the evidence rather than an entire case and lacks the jury deliberation that would usually occur before decisions were made. However, the main aim of this study was to understand if mock jurors are sensitive to the advantages of multiple experts independently assessing the evidence presented to them. The current study is the first to address this question and it is worth understanding the effect or lack thereof of these independent variables in a simple format before applying them to more in-depth trial simulation studies.

It is possible that the presentation of the variables in this study could have influenced the fact that no significant effect was found. An increase in the sample size did show a significant effect of crowd size on credibility ratings, but not for independence. Therefore, the presentation of independence in this study could have caused participants to not attend to that manipulation as much as crowd size. Despite attempts to clearly show independent versus non-independent manipulations through explanation and visualisation, this may not have been enough. As shown in figures 1a and b, it was explained that the verifying experts either independently arrived at the same conclusion or that they checked the work of the original expert and agreed with the conclusion. The graphic also showed experts grouped together for the non-independent group versus being separated for the independent groups. These might not have made it clear enough that the experts in the non-independent group knew the original experts' decisions before making their own decisions regarding the evidence. Further research could address this to see if this is the case by making the manipulation clearer.

**5.3 Future Research**

This thesis is the first step in understanding how the presentation of wisdom of the crowd decision-making, specifically multiple expert verifications and independence of their decisions, influences mock juror's credibility ratings of the evidence presented. The novelty of the current study allows for adaptations in a number of ways to extend the understanding of the possible uses of wisdom of the crowd decision-making. Blind verification is already used in forensic science, but results show that laypeople may not be inherently sensitive to the benefits of collective decision-making in this context. There are several reasons that this is important and should be addressed in future research. Initially, further research should explore the lack of effects found in the current research to test if this is due to limitations of the design or if mock jurors truly are not sensitive to crowd size and independence. In particular, the presentation of independence should be made clearer in future studies to test this theory.

If mock jurors are not sensitive to the independence of decisions as found in this study, future research could address why that is. Research has shown that mock jurors often use other indicators of expert credibility rather than the testimony itself which is problematic because it means that a seemingly highly credible expert could be more persuasive despite having a less credible testimony (Parrott et al., 2015). For example, based on the results of the current study, less accurate forensic science testimony (non-independent decisions) is rated similarly in credibility to more accurate testimony (independent decisions). Therefore, future research could investigate what might increase sensitivity to group size and independence. Salerno et al., (2017) found that cross-examination that advised jurors to focus on central information of the testimony helped them accurately evaluate the strength of the testimony. Further research could use cross-examination, education interventions, or judge's instructions to educate jury members on the potential dangers of non-independent or not verified decisions by experts. For example, Eastwood and Caldwell, (2015) used opposing expert witnesses to educate the jury about invalid forensic science testimony. Future studies could also use this to educate jurors on the effect that lack of independence has on decision-making processes to see if that helps them understand and interpret the wisdom of the crowd decision information. This could help ensure jurors are attending to the testimony rather than the expert and to assist with how this information is currently being used in court proceedings. Ultimately, to test and help with the application of this research, future studies would need to extend to real juries to reduce the limitations that come with mock juror studies.

Currently, despite the amount of research looking at jurors understanding and interpretation of expert witness testimony, there is still a wide range of measures used to test this. Some research focusses on the persuasiveness of the expert as rated by mock jurors (Martire et al., 2020), while others directly ask real jurors the believability or credibility of the expert (Champagne et al., 1991; Shuman et al., 1996). Brodsky et al., (2010) created the first published measure of expert witness credibility, which consisted of 20 items on four factors: confidence, likeability, trustworthiness, and knowledge. In contrast, some researchers prefer to test the outcomes of jurors' beliefs of credibility such as verdicts and sentencing. This is important because they are not always reflective of each other, for example, Cooper & Hall (2000) found that belief about the expert potentially being biased is not necessarily reflected in credibility ratings. Brodsky et al., (2009) also found that expert likeability did affect perceived trustworthiness but not sentencing decisions in his study. Therefore, it would be ideal to follow up the current research with some of these additional outcome variables to see if credibility ratings are affected in the same way by wisdom of the crowd decision-making information.

Prior beliefs and assumptions can influence the decision-making process by shaping how people interpret and believe information (Greenhoot et al., 2004). While the CSI effect has been shown to not affect jurors' beliefs as much as initially thought, jurors do hold prior beliefs and biases that can affect their opinions about evidence and verdicts (Schuller et al., 1994). Smith and Bull (2010) created the Forensic Evidence Evaluation Bias Scale to measure this bias. The current study did not measure prior beliefs to avoid any unintended influence of those questions, however, understanding the measuring prior beliefs could give more of an insight on how our independent variables affect belief change rather than just credibility ratings. For the current study, to test laypeople's general understanding of wisdom of the crowd information, we decided to leave the natural variation in sampling assumptions for the participants, however, future research could use something like the Forensic Evidence Evaluation Bias Scale to measure this.

This project addressed the influence of independence on interpreting wisdom of the crowd decision information, which is only one of the characteristics important for an accurate crowd, future research could also look at the influence of the others on credibility ratings. For example, previous research into the wisdom of the crowd effect has shown that diversity in the cognitive strategies, information use, and/or experience by individual experts should lead to larger effects in the accuracy of aggregated decisions (Balsdon et al., 2018). It would be interesting to know if mock jurors are aware of how this helps accuracy, especially given that they were not sensitive to this for the independence of decisions.

**5.4 Conclusions**

Results from the current study contradict predictions that mock jurors would be sensitive to the benefits of crowd size and independence in wisdom of the crowd decision-making. No significant effect was found for independence in either of the samples or the combined data analysis, apparent by the similarity between the mean credibility ratings between the two groups. The significance of this lack of an effect is important because jurors already interpret verified information without understanding how independence can influence its accuracy. While no significant effect was found for crowd size for the independent samples, the combined sample did show a significant quadratic effect. This indicated that there may be some level of sensitivity to crowd size among mock jurors. Further research is required to clarify this relationship and to address the lack of sensitivity to independence in practice.

**References**

Balsdon, T., Summersby, S., Kemp, R., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications, 3*(1), 1–13. https://doi.org/10.1186/s41235-018-0114-7

Baskin, D. R., & Sommers, I. B. (2010) Crime-Show-Viewing Habits and Public Attitudes Toward Forensic Evidence: The "CSI Effect" Revisited, *Justice System Journal, 31*(1), 97-113, DOI: 10.1080/0098261X.2010.10767956

Behson, S. J., & Koppl, R. (2013). Using Procedural Justice to Understand, Explain, and Prevent Decision-Making Errors in Forensic Sciences. *Organization Management Journal, 10*(2), 99–109. https://doi.org/10.1080/15416518.2013.801743

Blackwell, S., & Seymour, F. (2015). Expert Evidence and Jurors' Views on Expert Witnesses. *Psychiatry, Psychology, and Law, 22*(5), 673–681. https://doi.org/10.1080/13218719.2015.1063181

Bohner, G., Dykema-Engblade, A., Tindale, R., & Meisenhelder, H. (2008). Framing of Majority and Minority Source Information in Persuasion: When and How "Consensus Implies Correctness." *Social Psychology (Göttingen, Germany), 39*(2), 108–116. https://doi.org/10.1027/1864-9335.39.2.108

Bohner, G., Erb, H.-P., Reinhard, M.-A., & Frank, E. (1996). Distinctiveness across topics in minority and majority influence: An attributional analysis and preliminary data. *British Journal of Social Psychology*, *35*, 27–46.

Bornstein, B. (1999). The Ecological Validity of Jury Simulations: Is the Jury Still Out? *Law and Human Behavior, 23*(1), 75–91. https://doi.org/10.1023/A:1022326807441

Bornstein, B. H., Golding, J. M., Neuschatz, J., Kimbrough, C., Reed, K., Magyarics, C., & Luecht, K. (2017). Mock Juror Sampling Issues in Jury Simulation Research: A Meta-Analysis. *Law and Human Behavior, 41*(1), 13–28. https://doi.org/10.1037/lhb0000223

Brewer, S. (1998). Scientific Expert Testimony and Intellectual Due Process. *The Yale Law Journal, 107*(6), 1535–1681. https://doi.org/10.2307/797336

Brodsky, S., Griffin, M., & Cramer, R. (2010). The Witness Credibility Scale: An outcome measure for expert witness research. *Behavioral Sciences & the Law, 28*(6), 892–907. https://doi.org/10.1002/bsl.917

Brodsky, S. L., Neal, T. M. S., Cramer, R. J., & Ziemke, M. H. (2009). Credibility in the Courtroom: How Likeable Should an Expert Witness Be? *The Journal of the American Academy of Psychiatry and the Law, 37*(4), 525–532.

Champagne, A., Shuman, D., & Whitaker, E. (1991). An empirical examination of the use of expert witnesses in american courts. *Jurimetrics Journal, 31*(4), 375-392.

Choo, T.-H. (1964). Communicator Credibility and Communication Discrepancy as Determinants of Opinion Change. *The Journal of Social Psychology, 64*(1), 65–76. https://doi.org/10.1080/00224545.1964.9919544

Cooper, J., & Hall, J. (2000). Reaction of mock jurors to testimony of a court appointed expert. *Behavioral Sciences & the Law, 18*(6), 719–729. https://doi.org/10.1002/bsl.414

Cramer, R. J., Brodsky, S. L., & DeCoster, J. (2009). Expert witness confidence and juror personality: Their impact on credibility and persuasion in the courtroom. *The Journal of the American Academy of Psychiatry and the Law, 37*(1), 63–74. http://jaapl.org/content/37/1/63.long

Darke, P. R., Chaiken, S., Bohner, G., Einwiller, S., Erb, H.-P., & Hazlewood, J. D. (1998). Accuracy Motivation, Consensus Information, and the Law of Large Numbers: Effects on Attitude Judgment in the Absence of Argumentation. *Personality & Social Psychology Bulletin, 24*(11), 1205–1215. https://doi.org/10.1177/01461672982411007

Dror, I. E. (2015). Cognitive neuroscience in forensic science: Understanding and utilizing the human element. Philosophical Transactions. *Biological Sciences, 370*(1674), 20140255–. https://doi.org/10.1098/rstb.2014.0255

Eastwood, J., & Caldwell, J. (2015). Educating Jurors about Forensic Evidence: Using an Expert Witness and Judicial Instructions to Mitigate the Impact of Invalid Forensic Science Testimony. Journal of Forensic Sciences, 60(6), 1523–1528. https://doi.org/10.1111/1556-4029.12832

ForsterLee, L., Horowitz, I., Athaide-Victor, E., & Brown, N. (2000). The bottom line: the effect of written expert witness statements on juror verdicts and information processing. *Law and human behavior*, *24*(2), 259–270. https://doi.org/10.1023/a:1005415104323

Greenhoot, A. F., Semb, G., Colombo, J., & Schreiber, T. (2004). Prior beliefs and methodological concepts in scientific reasoning. *Applied Cognitive Psychology, 18*(2), 203–221. https://doi.org/10.1002/acp.959

Galton, F. (1907). The ballot-box. *Nature*, *75*(1952), 509-510.

Himelein, M. J., Nietzel, M. T., & Dillehay, R. C. (1991). Effects of prior juror experience on jury sentencing. *Behavioral Sciences & the Law*, *9*(1), 97-106.

Hosseini, M., Moore, J., Almaliki, M., Shahri, A., Phalp, K., & Ali, R. (2015). Wisdom of the Crowd within enterprises: Practices and challenges. *Computer Networks, 90*, 121–132. https://doi.org/10.1016/j.comnet.2015.07.004

Innocence Project (n.d.). Overturning Wrongful Convictions Involving Misapplied Forensics. Retrieved September 2, 2021, from https://innocenceproject.org/overturning-wrongful-convictions-involving-flawed-forensics/.

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *The British Journal of Psychology, 109*(4), 724–735. https://doi.org/10.1111/bjop.12291

Kattan, M. W., O'Rourke, C., Yu, C., & Chagin, K. (2016). The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones. *Medical Decision Making, 36*(4), 536–540. https://doi.org/10.1177/0272989X15581615

Keller, S. R., & Wiener, R. L. (2011). What are we studying? Student jurors, community jurors, and construct validity. *Behavioral sciences & the law*, *29*(3), 376-394. https://doi.org/10.1002/bsl.971

Kelley, H.H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation 15*. 192–238 Lincoln: University of Nebraska Press.

Koehler, J. J., Schweitzer, N. J., Saks, M. J., & McQuiston, D. E. (2016). Science, Technology, or the Expert Witness: What Influences Jurors' Judgments About Forensic Science Testimony? *Psychology, Public Policy, and Law, 22*(4), 401–413. https://doi.org/10.1037/law0000103

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., & Wolf, M. (2021). Pooling decisions decreases variation in response bias and accuracy. *iScience*, *24*(7), 1-19. https://doi.org/10.1016/j.isci.2021.102740

Langenburg, G., Champod, C., & Wertheim, P. (2009). Testing for Potential Contextual Bias Effects During the Verification Stage of the ACE-V Methodology when Conducting Fingerprint Comparisons. *Journal of Forensic Sciences, 54*(3), 571–582. https://doi.org/10.1111/j.1556-4029.2009.01025.x

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

MacCoun, R.J. (1989). Experimental Research on Jury Decision-Making. *Science, 244*(4908), 1046–1050. https://doi.org/10.1126/science.244.4908.1046

McCarthy Wilcox, A., & NicDaeid, N. (2018). Jurors' perceptions of forensic science expert witnesses: Experience, qualifications, testimony style and credibility. *Forensic Science International, 291,* 100–108. https://doi.org/10.1016/j.forsciint.2018.07.030

Martire, K., Edmond, G., & Navarro, D. (2020). Exploring juror evaluations of expert opinions using the Expert Persuasion Expectancy framework. *Legal and Criminological Psychology, 25*(2), 90–110. https://doi.org/10.1111/lcrp.12165

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior, 37*(3), 197–207. https://doi.org/10.1037/lhb0000027

Martire, K. A., & Montgomery-Farrer, B. (2020). Judging experts: Australian magistrates' evaluations of expert opinion quality. *Psychiatry, Psychology, and Law, 27*(6), 950–962. https://doi.org/10.1080/13218719.2020.1751334

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

McCabe, J. G., Krauss, D. A., & Lieberman, J. D. (2010). Reality check: A comparison of college students and a community sample of mock jurors in a simulated sexual violent predator civil commitment. *Behavioral Sciences & the Law, 28*(6), 730–750. https://doi.org/10.1002/bsl.902

McPhail, C. (1989). Blumer's Theory of Collective Behavior: The Development of a Non-Symbolic Interaction Explanation. *Sociological Quarterly, 30*(3), 401–423. https://doi.org/10.1111/j.1533-8525.1989.tb01528.x

McPhail, C. (2006). The Crowd and Collective Behavior: Bringing Symbolic Interaction Back In. *Symbolic Interaction, 29*(4), 433–463. https://doi.org/10.1525/si.2006.29.4.433

Murr, A. E. (2011). "Wisdom of crowds"? A decentralised election forecasting model that uses citizens' local expectations. *Electoral Studies*, *30*(4), 771-783. https://doi.org/10.1016/j.electstud.2011.07.005

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour, 2*(2), 126–132. https://doi.org/10.1038/s41562-017-0273-4

National Academy of Sciences (NAS), National Research Council. (2009). *Committee on identifying the needs of the forensic science community: Strengthening forensic science in the United States: A path forward*. Washington, DC: The National Academies Press.

Parrott, C., Neal, T., Wilson, J., & Brodsky, S. (2015). Differences in expert witness knowledge: do mock jurors notice and does it matter? *The Journal of the American Academy of Psychiatry and the Law, 43*(1), 69–81. https://digitalcommons.unl.edu/publicpolicyfacpub/36/

Reznicek, M., Ruth, R. M., & Schilens, D. M. (2010). ACE-V and the scientific method. *Journal of Forensic Identification, 60*(1), 87-103. http://proxy.library.adelaide.edu.au/login?url=https://www.proquest.com/scholarly-journals/ace-v-scientific-method/docview/194807964/se-2?accountid=8203

Scott, A. (2007). Peer review and the relevance of science. *Futures: the Journal of Policy, Planning and Futures Studies, 39*(7), 827–845. https://doi.org/10.1016/j.futures.2006.12.009

Shelton, Donald E., (2008) The 'CSI Effect': Does it Really Exist? *National Institute of Justice Journal, 259*, SSRN: https://ssrn.com/abstract=1163231

Shuman, D. W., Champagne, A., & Whitaker, E. (1996). Juror Assessments of the Believability of Expert Witnesses: A Literature Review. *Jurimetrics (Chicago, Ill.), 36*(4), 371–382.

Smith, L. L., & Bull, R. (2012). Identifying and measuring juror pre-trial bias for forensic evidence: development and validation of the Forensic Evidence Evaluation Bias Scale. *Psychology, Crime & Law*, *18*(9), 797-815. https://doi.org/10.1080/1068316X.2011.561800

Speckels, C. (2011). Can ACE-V Be Validated? *Journal of Forensic Identification, 61*(3), 201–209. https://www.proquest.com/docview/867849851?pq-origsite=gscholar&fromopenview=true

Ribeiro, G., Tangen, J., & McKimmie, B. (2019). Beliefs about error rates and human judgment in forensic science. *Forensic Science International, 297*, 138–147. https://doi.org/10.1016/j.forsciint.2019.01.034

Salerno, J. M., Bottoms, B. L., & Peter-Hagene, L. C. (2017). Individual versus group decision making: Jurors' reliance on central and peripheral information to evaluate expert testimony. *PloS One, 12*(9), e0183580–e0183580. https://doi.org/10.1371/journal.pone.0183580

Schuller, R. A., Smith, V. L., & Olson, J. M. (1994). Jurors' Decisions in Trials of Battered Women Who Kill: The Role of Prior Beliefs and Expert Testimony. *Journal of Applied Social Psychology, 24*(4), 316–337. https://doi.org/10.1111/j.1559-1816.1994.tb00585.x

Sorensen, L. (Ed.). (2019). *Mob Rule or the Wisdom of the Crowd?*. Greenhaven Publishing LLC.

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group Performance Depends on the Majority Rule. *Psychological Science, 9*(6), 456–463. https://doi.org/10.1111/1467-9280.00085

Surowiecki, J. (2004). The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations (1st ed.). Doubleday.

Smith, R. L., Kannemeyer, M., Adams, E., Nguyen, V. V., Munshaw, R., & Burr, W. S. (2020). Comparing jury focus and comprehension of expert evidence between adversarial and court-appointed models in Canadian criminal court context. *Canadian Society of Forensic Science Journal, 53*(2), 43–70. https://doi.org/10.1080/00085030.2020.1748284

Tangen, J., Kent, K., & Searston, R. (2020). Collective intelligence in fingerprint analysis. Cognitive Research: *Principles and Implications, 5*(1), 23–23. https://doi.org/10.1186/s41235-020-00223-8

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying Fingerprint Expertise. *Psychological Science, 22*(8), 995–997. https://doi.org/10.1177/0956797611414729

Tindale, R., Talbot, M., & Martinez, R. (2013). Decision Making. In Levine, J., *Group Processes* (pp. 77–204). Psychology Press. from https://doi.org/10.4324/9780203869673-14.

The Joint Effect of Scientific Knowledge and Photographic Evidence on Expert Witness Credibility. (2019). *Proceedings Of The Human Factors And Ergonomics Society Annual Meeting*, *63*(1), 1440-1444. doi: 10.1177/1071181319631418

Vanderkolk, J., 2011. ACE-V Examination Method. In: E. Holder, Jr, L. Robinson and J. Laub, ed., *The Fingerprint Sourcebook*, 1st ed. Washington: U.S. Department of Justice Office of Justice Programs, pp. 12 -17. https://www.crime-scene-investigator.net/fingerprintsourcebkchp9.pdf

Vidmar, N. (2005). Expert evidence, the adversary system, and the jury. *American Journal of Public Health*, *95*(S1), S137-S143. https://doi.org/10.2105/AJPH.2004.044677

Werner, C. M., Strube, M. J., Cole, A. M., & Kagehiro, D. K. (1985). The impact of case characteristics and prior jury experience on jury verdicts 1. *Journal of Applied Social Psychology*, *15*(7), 409-427. https://doi.org/10.1111/j.1559-1816.1985.tb02262.x

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd Effects in Unfamiliar Face Matching. *Applied Cognitive Psychology, 27*(6), 769–777. https://doi.org/10.1002/acp.2971

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. . (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PloS One, 10*(8), e0134269–e0134269. https://doi.org/10.1371/journal.pone.0134269

**Appendix 1 – Information Sheet**

## INFORMATION SHEET

### The purpose of the study

The aim of this project is to understand what affects credibility ratings of expert forensic evidence. This research is being conducted by Emma Tiggemann at The University of Adelaide in Australia. Email: emma.tiggemann@adelaide.edu.au.

### What's involved?

You will read 4 brief case reports about forensic expert decisions. You will be asked to answer some questions about your belief regarding the credibility of the evidence presented to you in each scenario. You will then also be asked some general and demographic questions. This study will take approximately 7 **minutes**.

### Risks

Your participation will help us to better understand how credible mock-jurors rate forensic science. There are no anticipated risks of participating that are greater than that of everyday living. If, however, you should find any question invasive or offensive, you are free to withdraw from the study at any time during your participation.

### Confidentiality and security of data

The data will be analysed and reported in such a way that responses will not be able to be linked to any individual. All data will be collected anonymously, and coded and publicly released in a way that makes uniquely identifying individual participants as difficult as possible. However, we cannot guarantee that it will be impossible to identify individual participants. For questions that request potentially identifying demographic information, respondents will always have the option to not respond.

### Ethics approval

This study has been approved by the School of Psychology Human Research Ethics Sub-Committee at the University of Adelaide (approval number: 21/42). If you wish to speak with an independent person regarding concerns or a complain or your rights as a participant, please contact the Convenor of the School of Psychology Human Research Ethics Sub-Committee at: paul.delfabbro@adelaide.edu.au.

**Appendix 2 – Consent form**

## CONSENT FORM

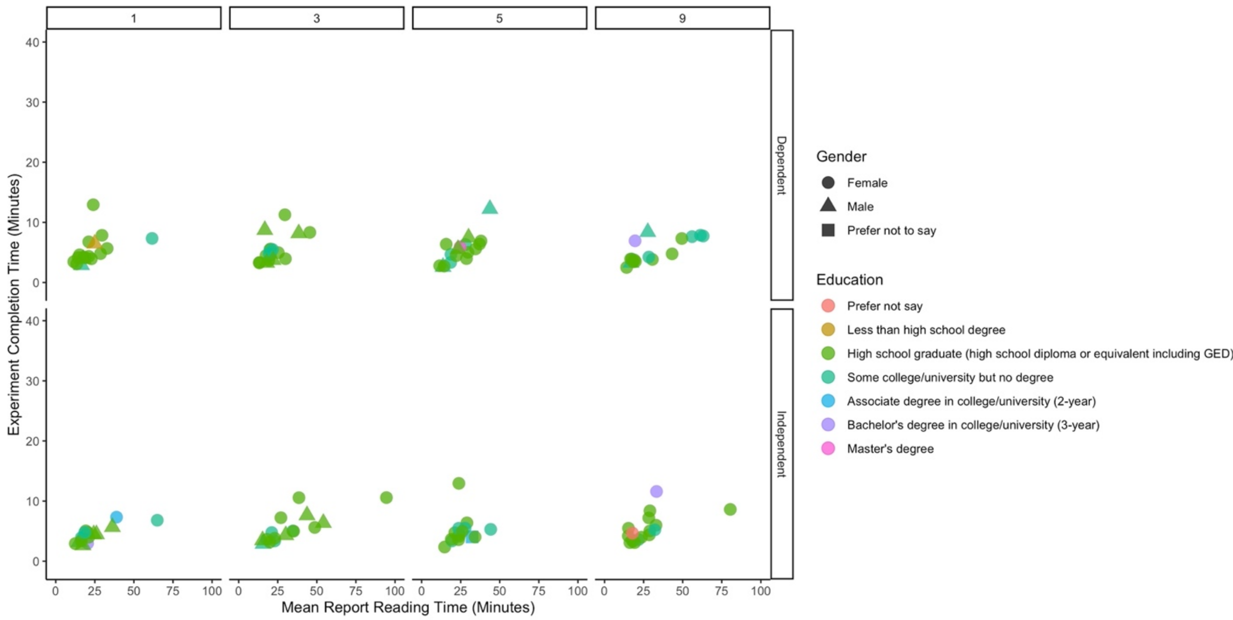
1. I have read and understood the information regarding the current study and freely consent to taking part in this survey.
2. I have had the project and the potential risks and burdens fully explained to my satisfaction.
3. I understand that I can withdraw at any time up until submission of my responses.
4. I understand that all information collected will be anonymous and coded and publicly released in a way that makes uniquely identifying individual participants as difficult as possible.
4. I understand that, if I choose to provide certain information, my data may not be completely non-identifiable.
5. I have been advised as to what data are being collected, what the purpose is, and what will be done with the data upon completion of the research.

6. I agree that research data gathered as part of the study may be published provided that neither my name, nor other identifying information, is used.
7. I agree that the de-identified research data gathered as part of the may be stored publicly and reused in other research projects.

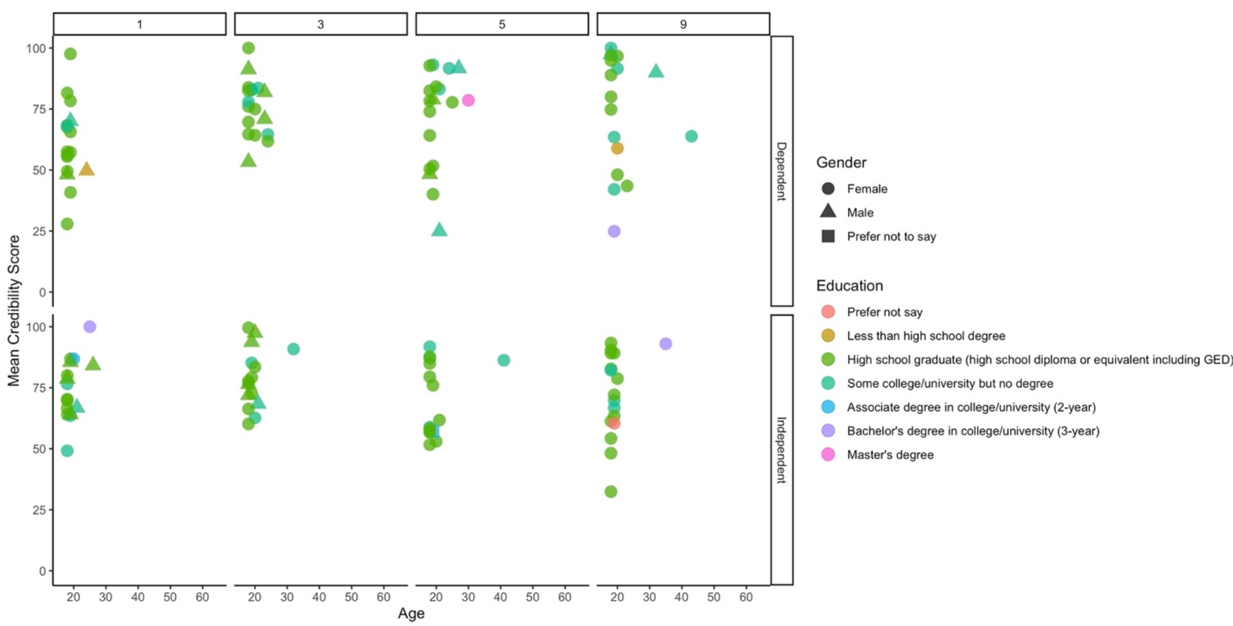
**Please indicate whether or not you are willing to participate in the study. If you choose to participate in this study, press "next".**

**Appendix 3 – Experiment 1a Demographics**

Experiment completion time (minutes) by report reading time (seconds)

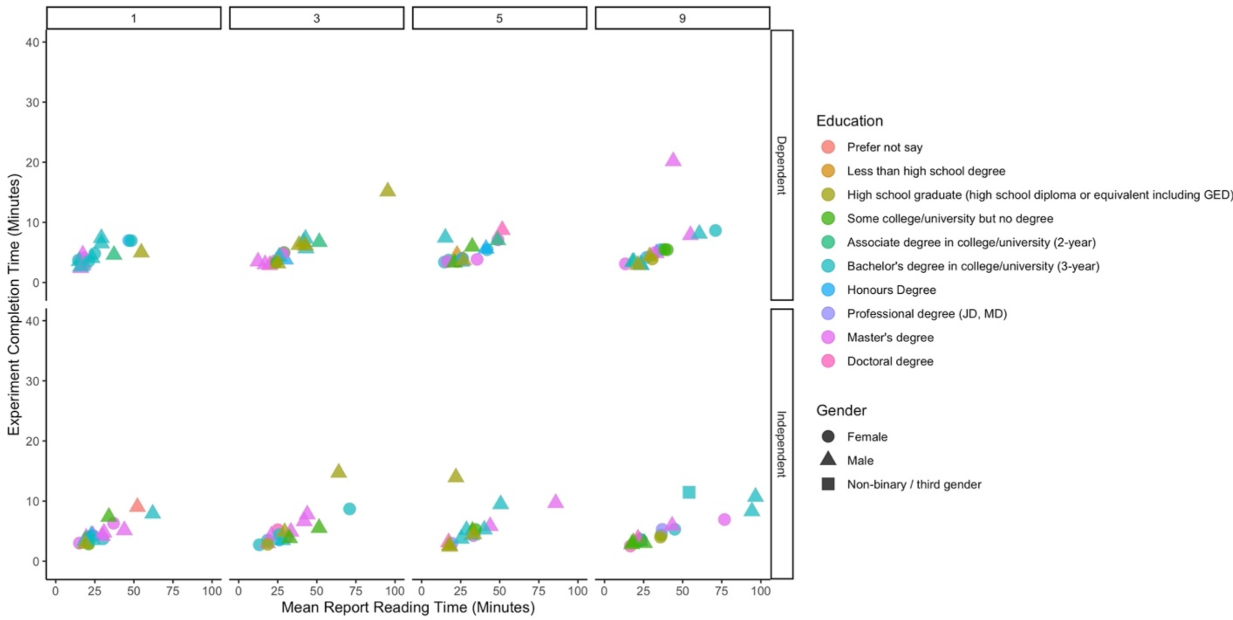



Experiment 1a – Credibility score (out of 100) by participant age

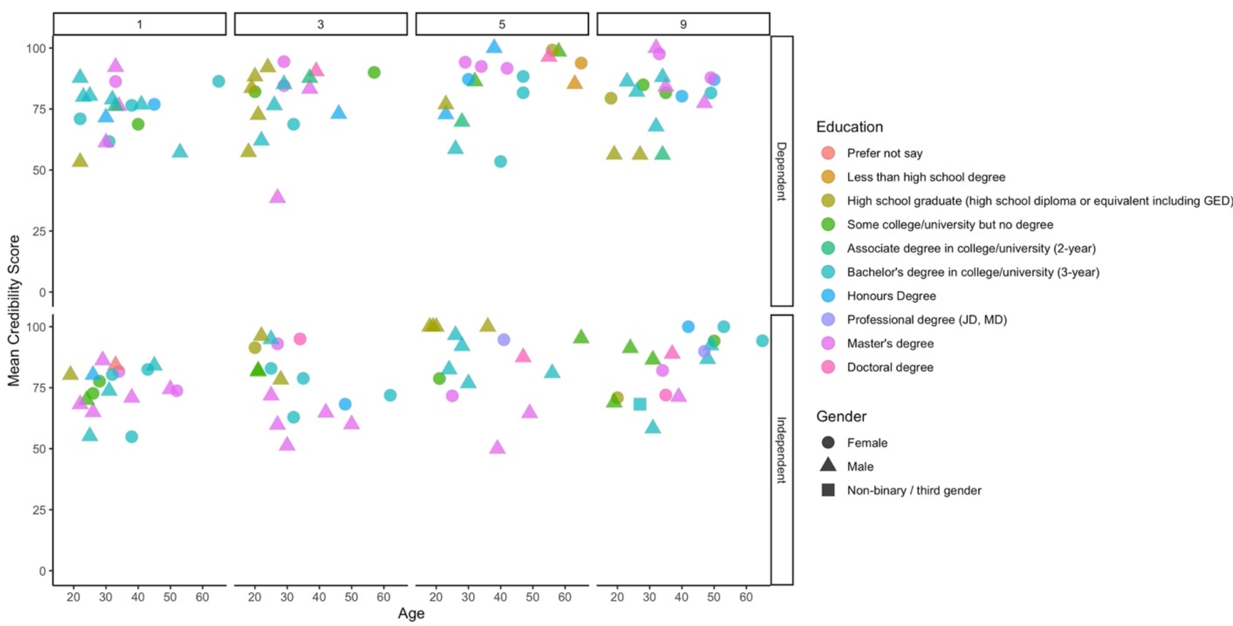


## Appendix 4 – Experiment 1b Demographics

Experiment completion time (minutes) by report reading time (seconds)

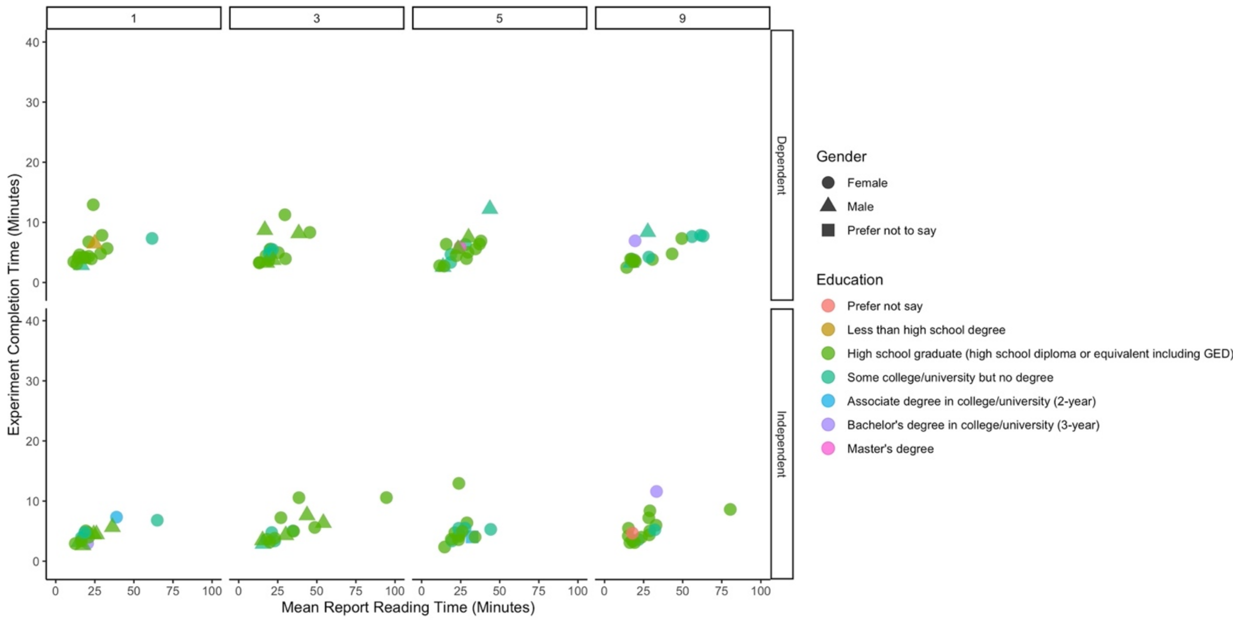



Credibility score (out of 100) by participant age

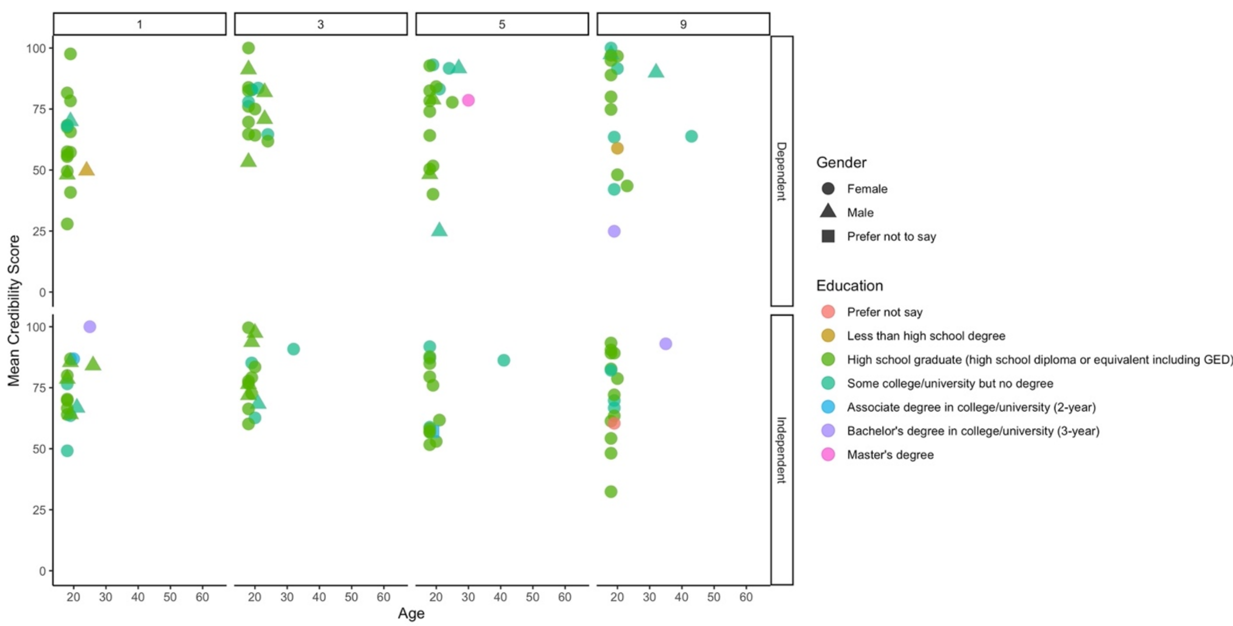


**Appendix 5 – Meta-analysis Demographics**

Experiment completion time (minutes) by report reading time (seconds)




Credibility score (out of 100) by participant age



## Appendix 6- Experiment 1a Data Analysis

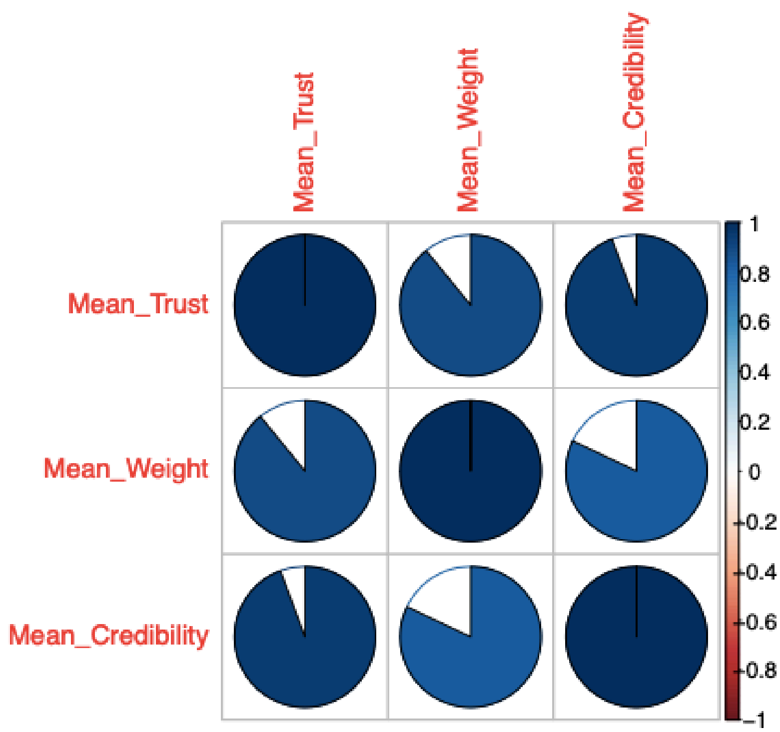
### Correlations

```
my_experiment %>%
  select(Mean_Trust,Mean_Weight,Mean_Credibility) -> Correlations

Correlations %>% cor() -> cor.matrix
cor.matrix
```

```
##                  Mean_Trust Mean_Weight Mean_Credibility
## Mean_Trust        1.0000000   0.8909947        0.9435282
## Mean_Weight       0.8909947   1.0000000        0.8201840
## Mean_Credibility  0.9435282   0.8201840        1.0000000
```

```
corrplot(cor.matrix, method="pie")
```



```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Credibility, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Credibility
## t = 33.213, df = 136, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9217548 0.9593709
## sample estimates:
##       cor
## 0.9435282
```

```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Weight
## t = 22.886, df = 136, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8505240 0.9209759
## sample estimates:
##       cor
## 0.8909947
```

```
cor.test(Correlations$Mean_Credibility, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Credibility and Correlations$Mean_Weight
## t = 16.719, df = 136, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7568042 0.8682846
## sample estimates:
##      cor
## 0.820184
```

## Internal Consistency

Reliability analysis is done with the alpha() function, which is found in the psych package.

```
credibility_score <- my_experiment %>%
  select(Mean_Trust, Mean_Weight, Mean_Credibility)
alpha(credibility_score)
```

```
##
## Reliability analysis
## Call: alpha(x = credibility_score)
##
##   raw_alpha std.alpha G6(smc) average_r S/N    ase mean sd median_r
##       0.96      0.96    0.95      0.88  23 0.0067   72 17     0.89
##
##  lower alpha upper     95% confidence boundaries
## 0.94 0.96 0.97
##
##  Reliability if an item is dropped:
##                  raw_alpha std.alpha G6(smc) average_r  S/N alpha se var.r
## Mean_Trust            0.90      0.90    0.82      0.82  9.1   0.0170    NA
## Mean_Weight           0.97      0.97    0.94      0.94 33.4   0.0050    NA
## Mean_Credibility      0.94      0.94    0.89      0.89 16.3   0.0099    NA
##                  med.r
## Mean_Trust        0.82
## Mean_Weight       0.94
## Mean_Credibility  0.89
##
##  Item statistics
##                    n raw.r std.r r.cor r.drop mean sd
## Mean_Trust       138  0.98  0.98  0.98   0.96   73 18
## Mean_Weight      138  0.94  0.94  0.89   0.87   68 19
## Mean_Credibility 138  0.96  0.96  0.94   0.90   75 17
```

### Assumptions

```
## Normality of Residuals Assumption
lm <- lm(Credibility_Score ~ Crowd_Size * Decision_Type, my_experiment) # Fit model

model.metrics <- augment(lm) %>%
  select(-.hat, -.sigma, -.fitted)
head(model.metrics, 3) # Inspect model
```

```
## # A tibble: 3 x 6
##   Credibility_Score Crowd_Size Decision_Type .resid  .cooksd .std.resid
##               <dbl> <ord>      <fct>          <dbl>    <dbl>      <dbl>
## 1              74   5          Dependent       2.55 0.000182      0.158
## 2              58.8 5          Independent   -10.6  0.00362      -0.659
## 3              89.2 9          Independent    17.0  0.00863       1.05
```

```
shapiro_test(model.metrics$.resid)
```

```
## # A tibble: 1 x 3
##   variable             statistic p.value
##   <chr>                    <dbl>   <dbl>
## 1 model.metrics$.resid     0.976  0.0162
```

### Generalized Linear Model

Let's look at some of the main differences between groups on each variable of interest using GLM with Poisson distribution. . .

```
my_experiment <- my_experiment
contrasts<-contr.poly(4,c(1,3,5,9))
model_glm <- glm(Credibility_Score ~ Crowd_Size * Decision_Type, family = Gamma, contrasts = list(Crowd
summary(model_glm)
```

```
##
## Call:
## glm(formula = Credibility_Score ~ Crowd_Size * Decision_Type,
##     family = Gamma, data = my_experiment, contrasts = list(Crowd_Size = contrasts))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.92112  -0.15754   0.01277   0.15612   0.51041
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.0142748  0.0003971  35.944   <2e-16
## Crowd_Size.L                        -0.0015654  0.0008148  -1.921   0.0569
## Crowd_Size.Q                         0.0014398  0.0008103   1.777   0.0779
## Crowd_Size.C                        -0.0013647  0.0007565  -1.804   0.0736
## Decision_TypeIndependent            -0.0006036  0.0005576  -1.082   0.2811
## Crowd_Size.L:Decision_TypeIndependent  0.0020798  0.0011283   1.843   0.0676
## Crowd_Size.Q:Decision_TypeIndependent -0.0017760  0.0011398  -1.558   0.1216
## Crowd_Size.C:Decision_TypeIndependent  0.0005489  0.0010766   0.510   0.6110
##
## (Intercept)                           ***
## Crowd_Size.L                          .
## Crowd_Size.Q                          .
## Crowd_Size.C                          .
## Decision_TypeIndependent
## Crowd_Size.L:Decision_TypeIndependent .
## Crowd_Size.Q:Decision_TypeIndependent
## Crowd_Size.C:Decision_TypeIndependent
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.05458358)
##
##     Null deviance: 9.0397  on 137  degrees of freedom
## Residual deviance: 8.3874  on 130  degrees of freedom
## AIC: 1195.4
##
## Number of Fisher Scoring iterations: 4
```

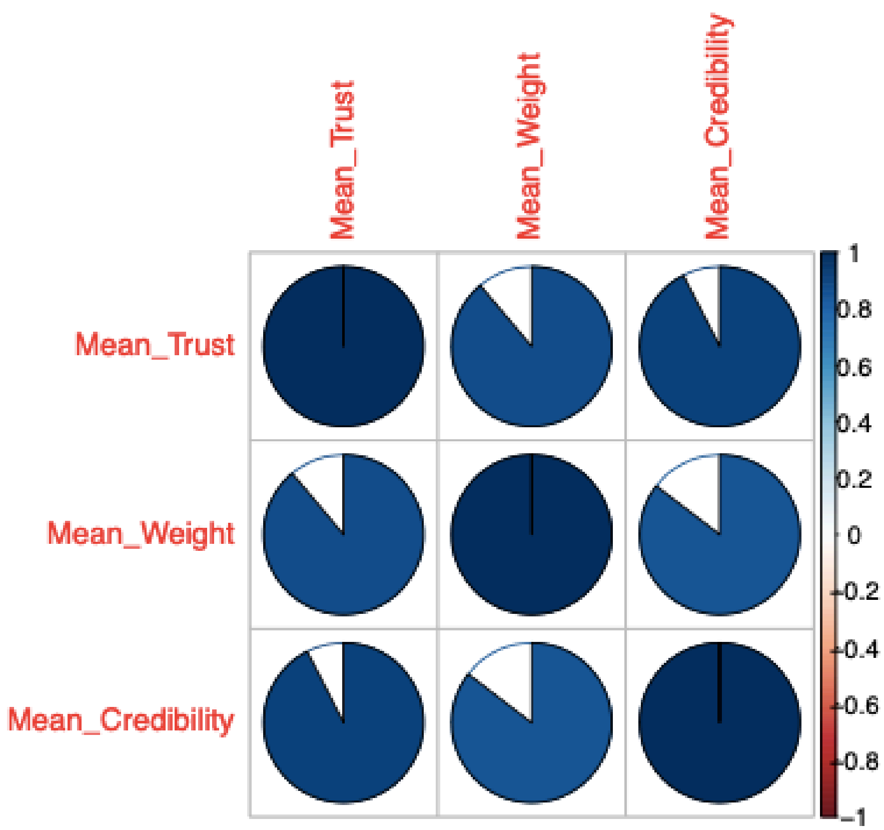
### Correlations

```
my_experiment %>%
  select(Mean_Trust,Mean_Weight,Mean_Credibility) -> Correlations

Correlations %>% cor() -> cor.matrix
cor.matrix
```

```
##                  Mean_Trust Mean_Weight Mean_Credibility
## Mean_Trust        1.0000000   0.8885705        0.9246111
## Mean_Weight       0.8885705   1.0000000        0.8538808
## Mean_Credibility  0.9246111   0.8538808        1.0000000
```

```
corrplot(cor.matrix, method="pie")
```



```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Credibility, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Credibility
## t = 29.027, df = 143, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8967620 0.9451649
## sample estimates:
##       cor
## 0.9246111
```

```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Weight
## t = 23.163, df = 143, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8484561 0.9185345
## sample estimates:
##       cor
## 0.8885705
```

```
cor.test(Correlations$Mean_Credibility, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Credibility and Correlations$Mean_Weight
## t = 19.619, df = 143, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8025835 0.8926427
## sample estimates:
##       cor
## 0.8538808
```

## Internal Consistency

Reliability analysis is done with the alpha() function, which is found in the psych package.

```
credibility_score <- my_experiment %>%
  select(Mean_Trust, Mean_Weight, Mean_Credibility)
alpha(credibility_score)
```

```
##
## Reliability analysis
## Call: alpha(x = credibility_score)
##
##   raw_alpha std.alpha G6(smc) average_r S/N    ase mean sd median_r
##       0.96      0.96    0.95      0.89  24 0.0059   80 13     0.89
##
##  lower alpha upper     95% confidence boundaries
## 0.95 0.96 0.97
##
##  Reliability if an item is dropped:
##                  raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
## Mean_Trust            0.92      0.92    0.85      0.85  12   0.0131    NA  0.85
## Mean_Weight           0.96      0.96    0.92      0.92  25   0.0065    NA  0.92
## Mean_Credibility      0.94      0.94    0.89      0.89  16   0.0098    NA  0.89
##
##  Item statistics
##                    n raw.r std.r r.cor r.drop mean sd
## Mean_Trust       145  0.97  0.97  0.96   0.94   80 14
## Mean_Weight      145  0.95  0.95  0.90   0.89   79 14
## Mean_Credibility 145  0.96  0.96  0.94   0.91   80 13
```

### Assumptions

```
## Normality of Residuals Assumption
lm <- lm(Credibility_Score ~ Crowd_Size * Decision_Type, my_experiment) # Fit model

model.metrics <- augment(lm) %>%
  select(-.hat, -.sigma, -.fitted)
head(model.metrics, 3) # Inspect model
```

```
## # A tibble: 3 x 6
##   Credibility_Score Crowd_Size Decision_Type .resid   .cooksd .std.resid
##               <dbl> <ord>      <fct>          <dbl>     <dbl>      <dbl>
## 1              78.9 3          Dependent      0.513 0.0000120     0.0416
## 2              91.2 5          Independent    5.21  0.00140       0.423
## 3              95.2 5          Independent    9.21  0.00437       0.748
```

```
shapiro_test(model.metrics$.resid)
```

```
## # A tibble: 1 x 3
##   variable             statistic   p.value
##   <chr>                    <dbl>     <dbl>
## 1 model.metrics$.resid     0.948 0.0000284
```

## Generalized Linear Model

Let's look at some of the main differences between groups on each variable of interest using GLM with Poisson distribution...

```
my_experiment <- my_experiment
contrasts<-contr.poly(4,c(1,3,5,9))
model_glm <- glm(Credibility_Score ~ Crowd_Size * Decision_Type, family = Gamma, contrasts = list(Crowd
summary(model_glm)
```

```
##
## Call:
## glm(formula = Credibility_Score ~ Crowd_Size * Decision_Type,
##     family = Gamma, data = my_experiment, contrasts = list(Crowd_Size = contrasts))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.63429  -0.08724   0.02924   0.10328   0.23558
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.0126189  0.0002326  54.260   <2e-16
## Crowd_Size.L                       -0.0006143  0.0004745  -1.295    0.198
## Crowd_Size.Q                        0.0008916  0.0004625   1.928    0.056
## Crowd_Size.C                        0.0003504  0.0004582   0.765    0.446
## Decision_TypeIndependent           -0.0001088  0.0003309  -0.329    0.743
## Crowd_Size.L:Decision_TypeIndependent -0.0005023  0.0006695  -0.750    0.454
## Crowd_Size.Q:Decision_TypeIndependent -0.0001669  0.0006570  -0.254    0.800
## Crowd_Size.C:Decision_TypeIndependent  0.0002192  0.0006586   0.333    0.740
##
## (Intercept)                        ***
## Crowd_Size.L
## Crowd_Size.Q                       .
## Crowd_Size.C
## Decision_TypeIndependent
## Crowd_Size.L:Decision_TypeIndependent
## Crowd_Size.Q:Decision_TypeIndependent
## Crowd_Size.C:Decision_TypeIndependent
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.02511366)
##
##     Null deviance: 4.2827  on 144  degrees of freedom
## Residual deviance: 3.8883  on 137  degrees of freedom
## AIC: 1170.7
##
## Number of Fisher Scoring iterations: 4
```

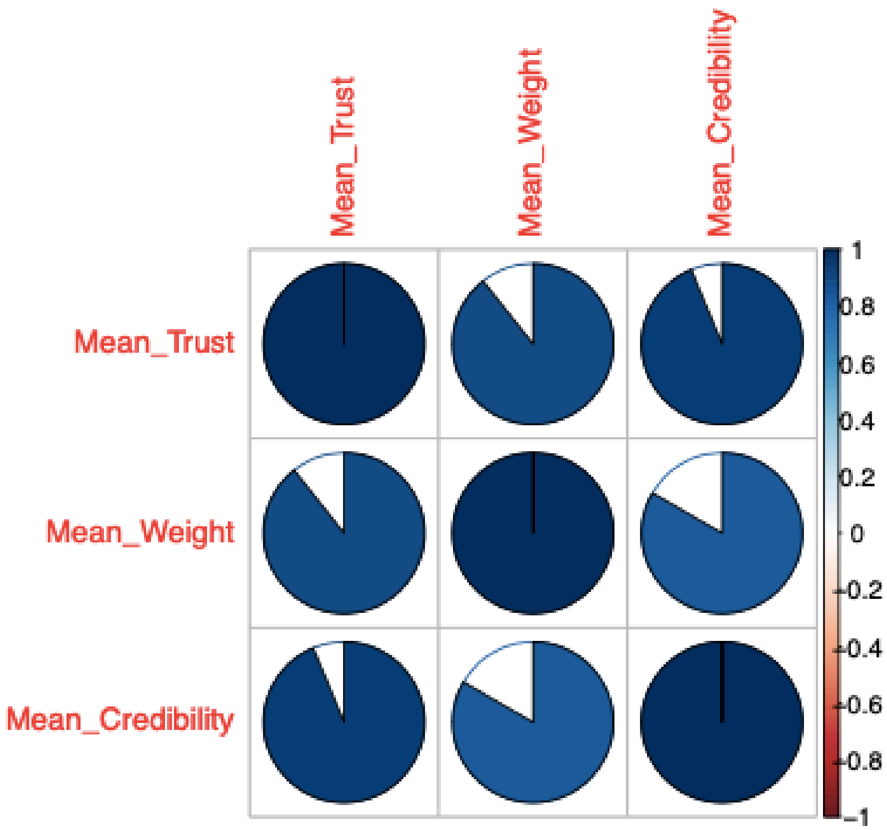
## Correlations

```
my_experiment %>%
  select(Mean_Trust,Mean_Weight,Mean_Credibility) -> Correlations

Correlations %>% cor() -> cor.matrix
cor.matrix
```

```
##                  Mean_Trust Mean_Weight Mean_Credibility
## Mean_Trust        1.0000000   0.8932556        0.9375049
## Mean_Weight       0.8932556   1.0000000        0.8331924
## Mean_Credibility  0.9375049   0.8331924        1.0000000
```

```
corrplot(cor.matrix, method="pie")
```



```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Credibility, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Credibility
## t = 45.163, df = 281, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9216541 0.9502319
## sample estimates:
##       cor
## 0.9375049
```

```
cor.test(Correlations$Mean_Trust, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Trust and Correlations$Mean_Weight
## t = 33.308, df = 281, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8669524 0.9145966
## sample estimates:
##       cor
## 0.8932556
```

```
cor.test(Correlations$Mean_Credibility, Correlations$Mean_Weight, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Correlations$Mean_Credibility and Correlations$Mean_Weight
## t = 25.257, df = 281, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7937015 0.8656897
## sample estimates:
##       cor
## 0.8331924
```

## Internal Consistency

Reliability analysis is done with the alpha() function, which is found in the psych package.

```
credibility_score <- my_experiment %>%
  select(Mean_Trust, Mean_Weight, Mean_Credibility)
alpha(credibility_score)
```

```
## Reliability analysis
## Call: alpha(x = credibility_score)
##
##   raw_alpha std.alpha G6(smc) average_r S/N    ase mean sd median_r
##       0.96      0.96    0.95      0.89  24 0.0045   76  16     0.89
##
##  lower alpha upper     95% confidence boundaries
## 0.95 0.96 0.97
##
##  Reliability if an item is dropped:
##                  raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
## Mean_Trust            0.90      0.91    0.83      0.83  10   0.0110    NA  0.83
## Mean_Weight           0.97      0.97    0.94      0.94  30   0.0039    NA  0.94
## Mean_Credibility      0.94      0.94    0.89      0.89  17   0.0068    NA  0.89
##
##  Item statistics
##                    n raw.r std.r r.cor r.drop mean sd
## Mean_Trust       283  0.98  0.98  0.98   0.95   77 16
## Mean_Weight      283  0.95  0.94  0.90   0.88   73 17
## Mean_Credibility 283  0.96  0.96  0.94   0.91   78 15
```

## Assumptions

```
## Normality of Residuals Assumption
lm <- lm(Credibility_Score ~ Crowd_Size * Decision_Type, my_experiment) # Fit model

model.metrics <- augment(lm) %>%
  select(-.hat, -.sigma, -.fitted)
head(model.metrics, 3) # Inspect model
```

```
## # A tibble: 3 x 6
##   Credibility_Score Crowd_Size Decision_Type .resid  .cooksd .std.resid
##               <dbl> <ord>      <fct>          <dbl>    <dbl>      <dbl>
## 1              74    5          Dependent      -4.13 0.000265     -0.272
## 2              58.8 5          Independent   -19.2  0.00626      -1.27
## 3              89.2 9          Independent    11.4  0.00215       0.754
```

```
shapiro_test(model.metrics$.resid)
```

```
## # A tibble: 1 x 3
##   variable             statistic    p.value
##   <chr>                    <dbl>      <dbl>
## 1 model.metrics$.resid     0.963 0.00000121
```

## Generalized Linear Model

Let's look at some of the main differences between groups on each variable of interest using GLM with Poisson distribution. . .

```
my_experiment <- my_experiment
contrasts<-contr.poly(4,c(1,3,5,9))
model_glm <- glm(Credibility_Score ~ Crowd_Size * Decision_Type, family = Gamma, contrasts = list(Crowd
summary(model_glm)
```

```
##
## Call:
## glm(formula = Credibility_Score ~ Crowd_Size * Decision_Type,
##     family = Gamma, data = my_experiment, contrasts = list(Crowd_Size = contrasts))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.96099  -0.14505   0.03801   0.14240   0.38394
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            0.0133646  0.0002250  59.386   <2e-16
## Crowd_Size.L                          -0.0010063  0.0004602  -2.186   0.0296
## Crowd_Size.Q                           0.0011070  0.0004531   2.443   0.0152
## Crowd_Size.C                          -0.0004079  0.0004367  -0.934   0.3511
## Decision_TypeIndependent              -0.0003376  0.0003180  -1.062   0.2893
## Crowd_Size.L:Decision_TypeIndependent  0.0006391  0.0006441   0.992   0.3220
## Crowd_Size.Q:Decision_TypeIndependent -0.0008169  0.0006401  -1.276   0.2029
## Crowd_Size.C:Decision_TypeIndependent  0.0003868  0.0006235   0.620   0.5355
##
## (Intercept)                           ***
## Crowd_Size.L                          *
## Crowd_Size.Q                          *
## Crowd_Size.C
## Decision_TypeIndependent
## Crowd_Size.L:Decision_TypeIndependent
## Crowd_Size.Q:Decision_TypeIndependent
## Crowd_Size.C:Decision_TypeIndependent
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.04101248)
##
##     Null deviance: 14.067  on 282  degrees of freedom
## Residual deviance: 13.530  on 275  degrees of freedom
## AIC: 2399
##
## Number of Fisher Scoring iterations: 4
```