EXAMINING SOCIAL INFLUENCE’S EFFECT ON DECISION-MAKING AND BAYESIAN TRUTH SERUM

A Thesis
Presented to
The Academic Faculty

by

Justin Sukernek

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Psychology in the
School of Psychology

Georgia Institute of Technology
May 2022

EXAMINING SOCIAL INFLUENCE’S EFFECT ON DECISION-MAKING AND BAYESIAN TRUTH SERUM

Approved by:

Dr. Rick Thomas, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Christopher Hertzog
School of Psychology
Georgia Institute of Technology

Dr. Chris Wiese
School of Psychology
Georgia Institute of Technology

Date Approved: August 27, 2021

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Rick Thomas for his excellent guidance and mentorship throughout my graduate training. I would also like to thank my committee, Dr. Chris Herzog and Dr. Chris Wiese, for supplying invaluable feedback and flexibility throughout this process.

I absolutely must thank Sophia Le for providing oft-needed support during a turbulent time in which nothing about the future was guaranteed. I am very lucky to have you in my life.

Finally, I would like to thank my parents, Warren and Linda, and my sister, Jenna, for whom I am nothing without. I am greatly indebted to you for what you have gifted me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

BTS	Bayesian truth serum
NFL	National Football League
ANOVA	analysis of variance
Sum Sq	sum of squares
Chi-Sq	chi-squared value
DF	degrees of freedom
LogLik	log-likelihood

SUMMARY

Decision-making—whether individual or in groups—can be subject to revision based on social influence, often pulling one's opinions towards the apparent consensus (Mason, Conrey, & Smith, 2007). Social influence has been shown to damage the effectiveness of wisdom of the crowd, suggesting that perhaps the crowd is wise—but only when the members do not interact with each other (Lorenz, Rauhut, Schweitzer, & Helbing, 2011). An interesting, unexplored method to study the effect of social influence would be to apply it to the Bayesian truth serum (BTS), a multi-faceted measure of judgment ability. In its pure application, the truth serum is both a measure of judgment and a way to increase truth-telling and information quality, but currently it is unclear if social influence may have a positive or negative effect on the serum's effectiveness (Frank, Cebrian, Pickard, & Rahwan, 2017). I conduct a multi-experiment study to elucidate further the possible adverse effects of social influence, and test Bayesian truth serum's robustness when combined with the influence of others' opinions. In combination, the five experiments show evidence for social influence disinforming participants; this disinformation effect appears to be detrimental to the Bayesian truth serum. Finally, these experiments cast doubt on the Bayesian truth serum's predictive ability in several different task contexts. Additionally, in one experiment we find evidence that disagreeing with social influence improves reasoning ability. Overall, this study contributes to the social influence, disinformation, and BTS literatures.

CHAPTER 1. INTRODUCTION

The influence of information coming from other people gains significance as internet-driven interconnectedness rises. While researchers are currently investigating the capabilities of disinformation, particularly in social media (Shu, Wang, Lee, & Liu, 2020), it is imperative to fully understand the root of disinformation's power: the effects of social influence.

Social influence can appear in a variety of ways, with advice-taking as a primary method. When privy to others' opinions, subjects have demonstrated a willingness to revise their own opinions to align with their advisors (Sniezek & Buckley, 1995). However, when given multiple opinions, people will give more weight to the opinions that are closest to their own (Yaniv & Milyavsky, 2007). Regardless, people revise their opinions even when receiving conflicting advice. However, these studies use a paradigm that labels the additional advice as coming from advisors, meaning that some participants may weigh the "advisor's" opinion more highly. These studies also involve real people; when participants doubt the existence of their advisor, they may discount the advice. One key difference between the advice-taking literature and social influence as it appears in this study is that here, independence between participants is disrupted as the influence comes from within the crowd. In contrast, in advice-taking tasks, advisors are typically not also respondents.

The effects of social influence on decision-making can emerge through examining crowd wisdom, a statistical phenomenon that utilizes aggregates of independent decision-makers to produce accurate results. Crowd wisdom has been shown to outperform experts in various sectors, such as trading stocks and even medical diagnosis (Nofer & Hinz, 2014; Wolf, Krause, Carney, Bogart, & Kurvers, 2015). When exposed to peer estimates within a homogenous partisan network, the accuracy of crowd wisdom improved (Becker, Porter, & Centola, 2019). This finding is particularly notable because elsewhere, more independence between individuals improved performance (Nofer & Hinz, 2014).

In line with Nofer & Hinz, another study found that social influence can be damaging: research using a statistic estimation task found that the accuracy of crowd wisdom decreased when participants were exposed to aggregated and non-aggregated peer estimates (Lorenz et al., 2011). This effect is theorized to be the result of three main reasons. First, social influence reduces the diversity of opinions without increasing accuracy. Second, the reduced diversity leads to a smaller range of answers that may not even capture the correct answer. Finally, people report increased confidence due to social influence, creating a distorted view of accuracy that is especially dangerous with high-difficulty questions (Lorenz et al., 2011).

Well-known effects also demonstrate the possible negatives of social influence. Frequently, subjects may follow their peers' decision-making despite being faulty—also known as herd behavior (Banerjee, 1992). Classically, these effects have been observed in studies of conformity, but the experiments involved in this proposal involve far less pressure and influence than something like Solomon Asch's line judgment task (Asch, 1951).

Social influence effects are important to study since they allow us to learn more about the way people integrate others' opinions. This is particularly applicable when thinking about group work and collaboration. It is common for professionals across various domains to consult their peers before making a decision; for example, doctors discussing possible diagnoses. While some collaboration may be involved in this case, the opinions are formed after-the-fact via social influence. Outside of professional contexts, studying social influence—particularly through online experiments—can provide insight into internet-based information, disinformation, and misinformation's impact.

1.1 Bayesian Truth Serum

A novel way to measure judgment ability is to apply the Bayesian truth serum. Bayesian truth serum is a method to find the best judge of a population by finding the surprisingly common answer (Prelec, 2004). The truth serum is comprised of two components: the information score and the prediction score. The information score is calculated by comparing a respondent's answers to the population. The information score is also known as the "surprisingly popular" signal; participants will receive an increased score for submitting the response with the largest disparity between its selection probability (popularity) and population prediction. The prediction score is calculated by comparing a respondent's expected population distribution of how people answered the question to the actual distribution. Thus, people receive high BTS scores when they have a preference (choose) the surprisingly popular answer and accurately predict the population distribution of preference.

Prelec theorizes the truth serum has several additional properties and use cases, such as to induce truth telling and help forecast the future. Alternatively, Bayesian truth serum has also been used to find correct answers in crowd wisdom tasks (Prelec, Seung, &

McCoy, 2017). Over seven experiments testing participants on verifiable questions ranging from state capitals to art value, Prelec and colleagues found that the surprisingly popular signal (BTS) limited error the most, compared to other methods such as consensus and weighting by confidence. Additionally, the researchers provide evidence that the BTS heavily outperforms the average respondent. Consequently, Prelec et al. conclude that the truth serum could be used as a powerful alternative to standard crowd wisdom.

Few experimental manipulations have been investigated with the BTS, as the current literature mainly focuses on testing the serum's properties with different samples and in different tasks (Witkowski & Parkes, 2012; Frank, Cebrian, Pickard, & Rahwan, 2017). This is somewhat surprising, as the truth serum is incredibly simple to implement: participants only need to answer a question, then their predicted distribution of how the population would answer the question (in most cases, the population in question is the population of participants). Even though we are using the truth serum as our primary measure, the experiments in the proposal are also testing the truth serum itself.

Many of the researchers testing the serum are interested in its theorized ability to accurately forecast. Lee, Danileiko, & Vi (2018) tested the truth serum's ability to predict winners of NFL games, and found that the surprisingly popular signal outperformed all other signals. Furthermore, the authors also found evidence that domain knowledge—in this case self-reported—may play a large role in the effectiveness of the signal. This conclusion is shared by Rutchick, Ross, Calvillo, & Mesick (2020), who used a more objective questionnaire to determine domain knowledge. Rutchick et al. hypothesized that expertise may matter more in forecasting tasks, as compared to general trivia items such as those presented in Prelec et al. (2017). Prelec himself recognized the potential integral role that domain knowledge plays in a forecasting context and helped devise a method that would heavily weight the more informed participants (Olsson, de Bruin, Galesic, & Prelec, 2019). Since the BTS is fairly new, these three studies reflect the entirety of the forecasting-related truth serum literature, necessitating further experimentation—such as study 2c in this experiment. Additionally, the studies cited in this paper comprise the entirety of the psychologically relevant BTS literature.

As outlined above, the comparison between predicted distributions of the population and the actual distribution plays a large role in the BTS scoring. Accordingly, it is likely that social influence may have a significant effect on truth serum outcomes. When predicting the distribution of the population of our participants, a subject's proposed distribution may skew towards the direction of the influence they receive. Therefore, there is a chance that social influence disrupts, or possibly even enhances, the accuracy of the truth serum.

CHAPTER 2. EXPERIMENT ONE

My main hypothesis for the first experiment is as follows:

- H1: The presence of social influence will have a significant effect on the individual Bayesian truth serum scores; the direction of the effect will be dependent on the amount of disagreement between the participant and the influencer; participants who disagree with their influence will have better truth serum scores; participants who agree will have worse scores.

Additional experimental hypotheses:

- H2: The main effect of pre-test and post-test vs post-test only will be statistically insignificant; thus, any changes in BTS for participants who provide their preference and population distribution twice will not be due to order effects.
- H3: Participants who receive low quality information will disagree with their influencer more than participants who receive high quality information.

2.1 Methodology

This experiment used a 3 x 2 x 2 between-subjects design, varying who participants receive influence from in the second phase (Profile A, Profile B, no one), the quality of information given to each participant (low or high), and if participants are given a pre-test as well as a post-test or only a post-test. We selected a crime scene investigation task for several reasons. First, a crime scene investigation paradigm is new to the social influence literature and the Bayesian truth serum literature; it is particularly important to investigate diverse tasks in the latter, as the performance of BTS in many tasks is still unknown. Second, a crime scene investigation involves processes like counterfactual reasoning and forecasting. These decision-making processes are important to study due to their significance in domains such as intelligence analysis.

2.1.1 Participants

We used the Qualtrics sampling system to gather participants. Participants received a small amount of money determined by Qualtrics for their participation. Participants submitted basic demographic information. In total, we collected data from 624 participants.

2.1.2 Materials

The experiment was fully conducted through a Qualtrics survey. First, a mock crime scene vignette was shown to the participants that described a crime scene and two possible suspects found fleeing the scene. After the vignette, participants saw a slew of pictures with varying relevance. For example, participants saw pictures of a singular bloody shoe, a table full of narcotics, and a bullet casing. An important part of the vignette is that there is a correct answer to the question of who committed the crime.

In the condition where participants were given more info, incriminating evidence like blood tests were given in addition to the other information. In theory, the additional evidence reduces vagueness and allows us to elucidate any moderating effect social influence may have on the decision-making process. We hypothesized that participants using the higher quality evidence logically should have an easier time making conclusions, leading to less disagreement. All pieces of evidence, ambiguous and diagnostic, are included in the appendix.

Participants in the social influence condition were given information from one of four profiles, depending on their assigned evidence quality condition. Two profiles supported answering Suspect A, while the other two profiles supported answering Suspect B. The profiles changed with evidence quality; for example, the Suspect B high-quality profile mentioned the additional forensic testing found in the high-quality condition. These two profiles were created after running a pilot study identical to the control condition, gathering real responses that participants submitted. The profiles did not disclose their predicted distribution of the population or their final answer, key parts of the truth serum calculation.

2.1.3 Procedure

The experiment progressed in three stages, with each stage lasting 5 minutes. The first stage was always an individual stage—participants read the given vignette and brainstormed hypotheses on who committed the crime and why. If a participant was in the pre-test condition, they were asked to answer who committed the crime, write down their hypotheses, and answer what percent of people in the population would select each option. Participants could select Suspect A, Suspect B, both, or neither as their answer.

The second stage was the influence stage. Participants read hypotheses from one of the influence profiles. Participants with no influencer skipped this stage. The third stage was identical to the first stage, with everyone generating hypotheses and receiving a post-test.

2.2 Results

BTS scores were calculated for each participant, and an additional agreement label was given to each participant depending on if their final judgment aligned with the influence they received. Participants who received no influence were labeled as such. Statistical models were conducted for each of the hypotheses. An ANOVA was conducted predicting post-test BTS scores with quality, pre-test, and influence terms (**Table 1**). The results were largely not statistically significant. The effect size of the difference in BTS scores between groups that received influence and groups that did not was 0.06, indicating a small effect.

To test H1, an initial ANOVA was conducted examining the main effect of social influence on post-test BTS scores (**Table 2**). While there was an observable difference in scores between the influence types (**Figure 1**), the main effect was not statistically significant. In the second part of H1, we predicted that the effect's direction would change depending on the agreement between the participant and the social influence. To test this hypothesis, we ran two additional models, examining the relationship between agreement and the BTS. The first model included influence and agreement as well as an interaction term (**Table 3**). While the main effect of influence was not significant, agreement and the interaction were highly significant. Similarly, the second model included only the agreement term and removed all participants who did not receive any influence (**Table 4**). This model's results also showed a highly significant effect of agreement.

Figures 2 and 3 also illustrate that the BTS scores and the BTS components depended on agreement level. The models and the figures support the second half of H1, despite the main effect of influence not being significant. I theorize that disagreement improved BTS scores due to the prediction score. When participants generate estimated population distributions for each answer, they typically overrate the popularity of the answer they choose. By receiving social influence contrary to their choice, participants incorporate that choice into their predicted distribution, pulling it closer to the actual distribution. We observe this increase in the higher prediction scores (Pscore) on Figure 3.

We tested H2 by conducting multiple statistical tests. First, a model examining the main effect of pre-test and post-test vs post-test only on post-test BTS scores (**Table 5**) yielded significant results. Second, two paired t-tests were conducted, one including and one excluding participants who received influence. Both t-tests were significant, with p-values of 0.00025 and 0.01348, respectively. **Figure 4** plots BTS scores by pre-test condition, showing that participants who received a pre-test had significantly lower BTS scores than those who received only the post-test. Finally, the main effect of pre-test was also included in the overall ANOVA as seen on **Table 1**. Additionally, the pre-test manipulation did not significantly affect agreement levels, as there were nearly equal amounts of agreement in each pre-test condition.

We tested H3 with a model predicting agreement using quality (**Table 6**). While the results were not statistically significant, **Figure 5** shows an observable difference in disagreement between participants who received low and high-quality information.

Two multinomial regression models were conducted to predict answer preference. The first included all the main terms: influence, pre-test, and evidence quality (**Table 7**). The second only included influence (**Table 8**). In the full model, influence and quality had significant effects on answer preference, as well as their interaction. In the model only including influence, influence’s effect was not statistically significant. We ran four separate ANOVA models examining all three main terms' effect on the four different distributional judgments (the predicted distributions for answers one, two, three, and four). Most terms were insignificant; however, the quality main effect was significant in each model. Additionally, the influence by quality interaction was significant for the models predicting people's distributional judgments for answers two (only suspect two) and three (both suspects).

The Bayesian truth serum poorly predicted truth in this context, as participants who chose the fourth option (neither) received the highest score, on average (**Figure 6**). The correct answer was the second option (only suspect 2), which received the third-worst truth serum score, on average.

Table 1. ANOVA conducted with all terms predicting BTS scores.

	DF	*Sum Sq*	*F-value*	*P-value*
Quality	1	93	16.020	p < 0.0001***
PreTest	1	69	11.835	0.0006***
Influence	2	9	0.751	0.4723
Q*PT	1	3	0.456	0.4996
Q*I	2	4	0.367	0.6931
PT*I	2	6	0.559	0.5722
Q*PT*I	2	19	1.669	0.1893
Residuals	612	3548	NA	NA

Table 2. ANOVA conducted examining the main effect of social influence with BTS scores as the dependent variable.

	DF	*Sum Sq*	*F-value*	*P-value*
Influence	2	9	0.723	0.486
Residuals	621	3742	NA	NA

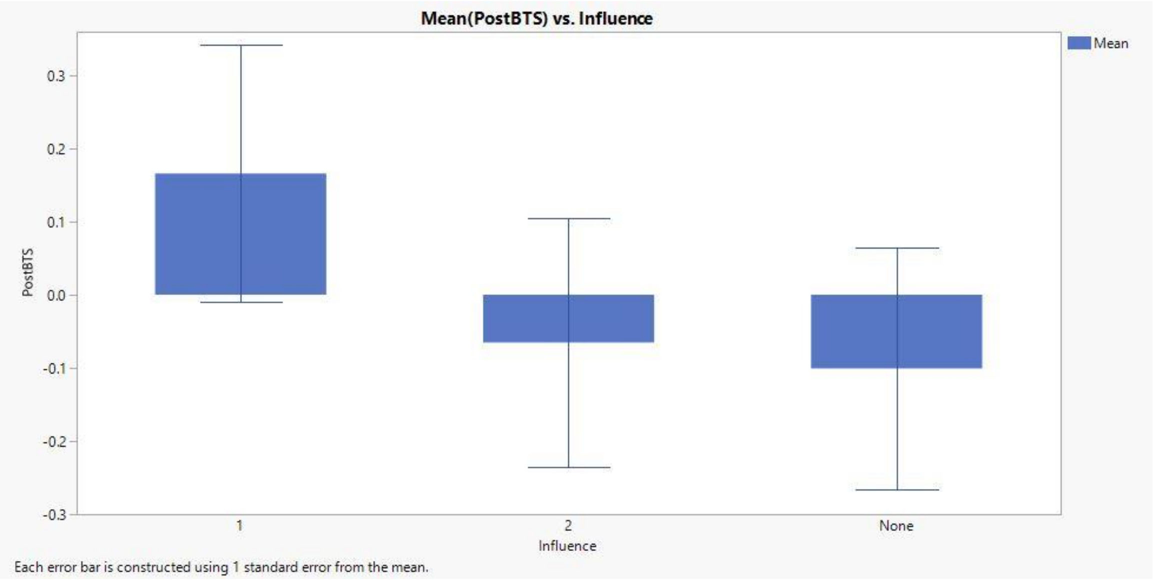


Figure 1. Plotting BTS scores and influence type.

Table 3. ANOVA conducted predicting BTS scores including influence and agreement as terms (participants receiving no influence removed).

	DF	*Sum Sq*	*F-value*	*P-value*
Influence	1	5.5	1.075	0.3
Agreement	1	357.8	69.323	< 0.0001***
Interaction	1	82.8	16.037	< 0.0001***
Residuals	412	2126.7	NA	NA

Table 4. ANOVA examining effect of Agreement on BTS scores (participants receiving no influence removed).

	DF	*Sum Sq*	*F-value*	*P-value*
Agreement	1	347	65.54	< 0.0001***
Residuals	414	2226	NA	NA

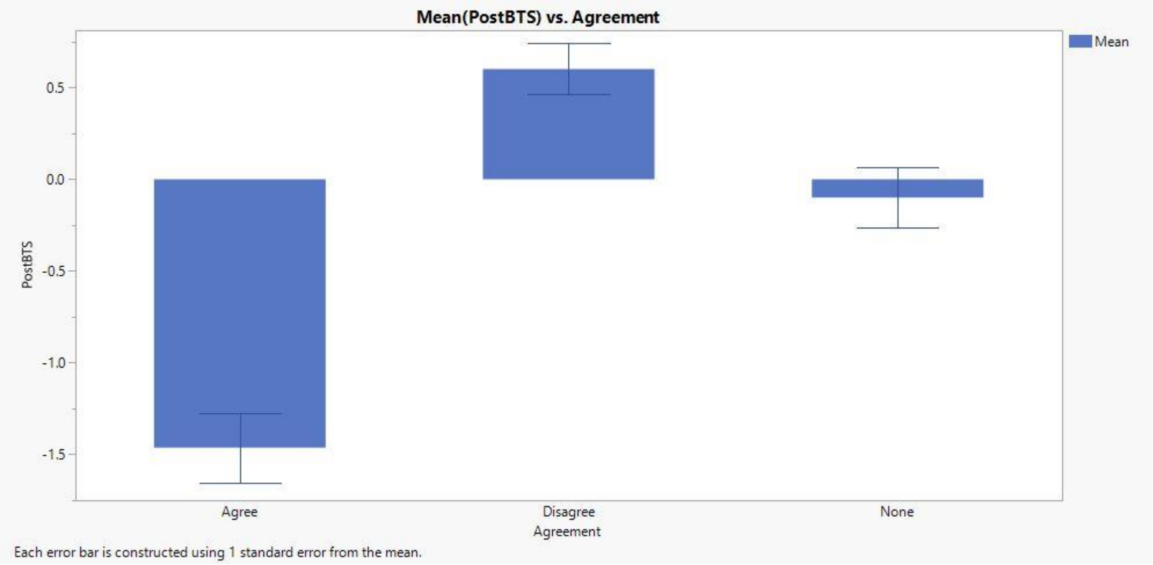


Figure 2. Plotting the BTS scores for participants based on agreement level.

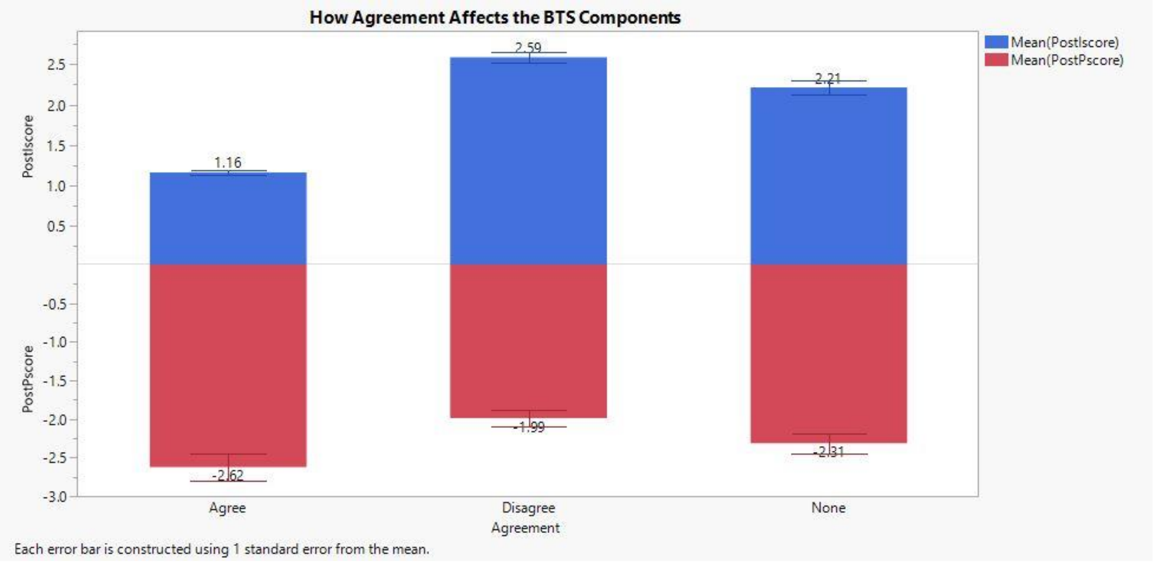


Figure 3. Plotting the BTS components for participants based on agreement level.

Table 5. ANOVA examining effect of pre-test condition on post-test BTS scores.

	DF	*Sum Sq*	*F-value*	*P-value*
PreTest	1	69	11.59	0.000706***

Residuals	622	3682	NA	NA

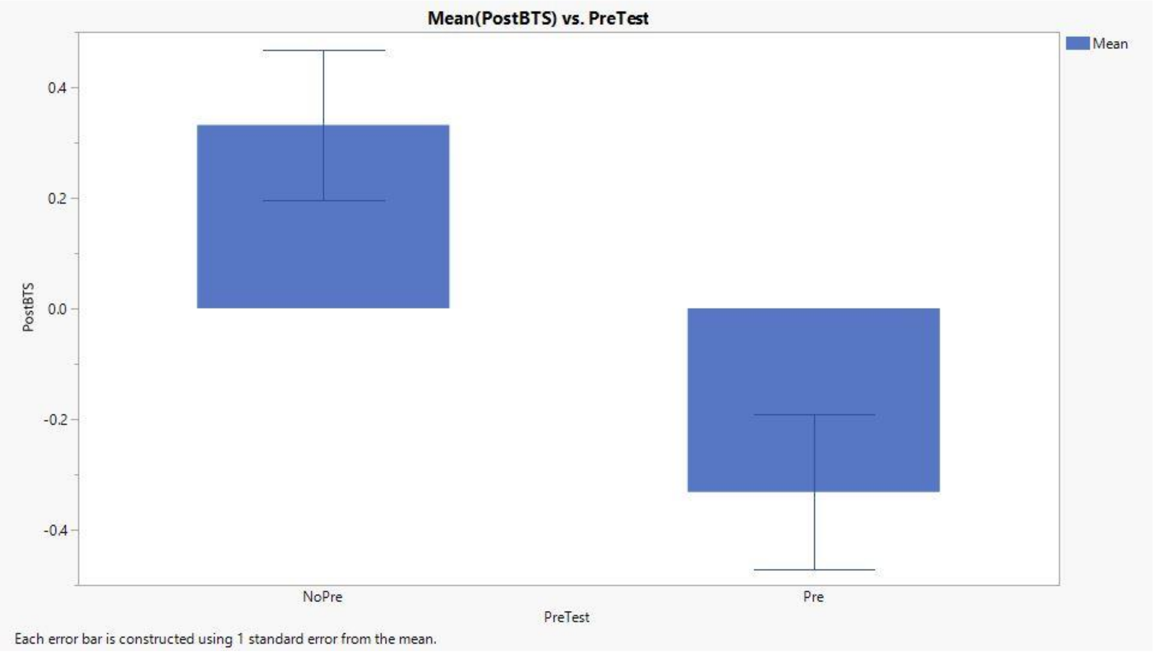


Figure 4. Plotting BTS scores by pre-test condition.

Table 6. ANOVA conducted predicting agreement with data quality (participants receiving no influence removed).

	DF	*Sum Sq*	*F-value*	*P-value*
Quality	1	0.54	2.77	0.0968
Residuals	414	80.84	NA	NA

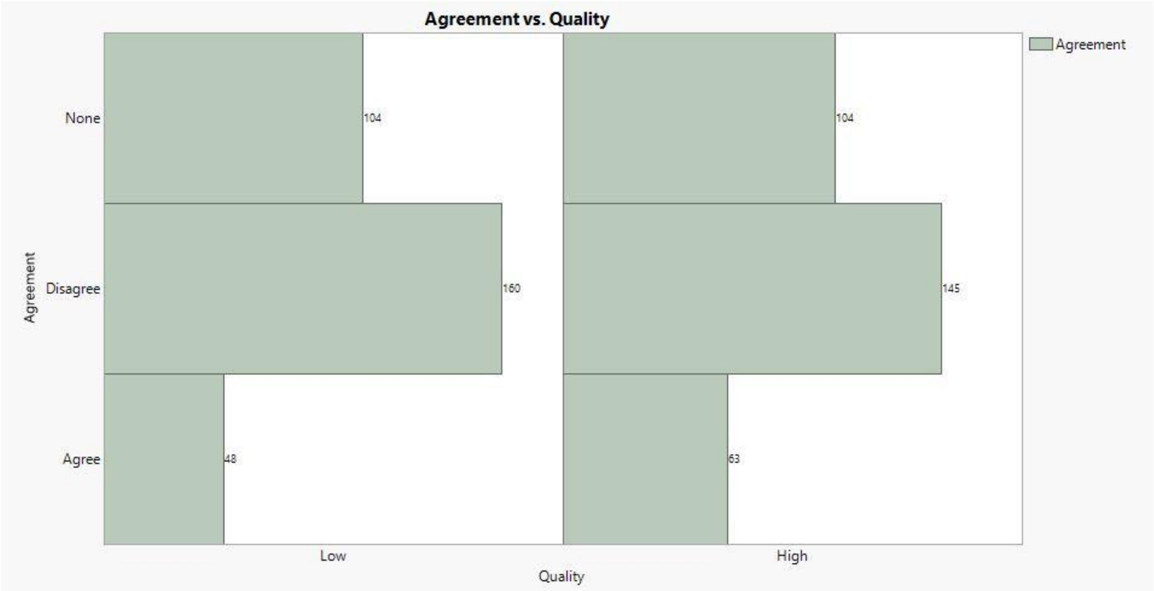


Figure 5. Plot of agreement counts per quality condition.

Table 7. Multinomial regression model predicting answer preference with influence, pre-test, and quality.

	ChiSq	*DF*	*P-value*
Influence	16.578	6	0.0109*
Quality	23.119	3	p < 0.0001***
PreTest	2.030	3	0.5662
I*Q	18.578	6	0.0049**
I*PT	10.294	6	0.1128
Q*PT	2.304	3	0.5117
I*Q*PT	10.121	6	0.1196

Table 8. Multinomial regression model predicting answer preference with only influence.

	ChiSq	*DF*	*P-value*
Influence	0.8560	6	0.9905

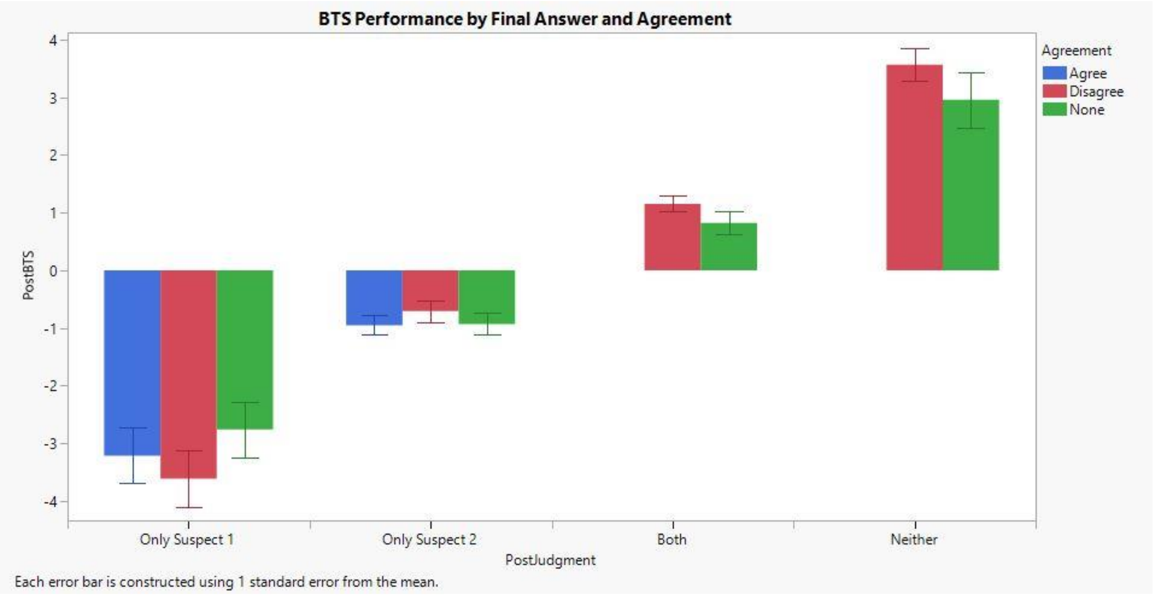


Figure 6. BTS scores by answer and agreement.

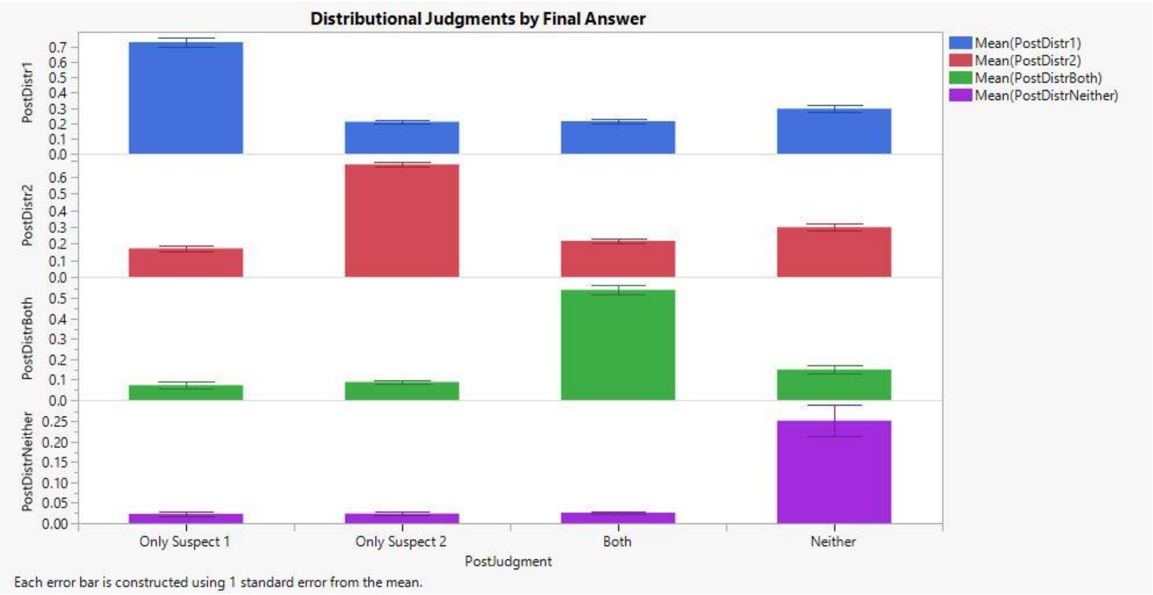
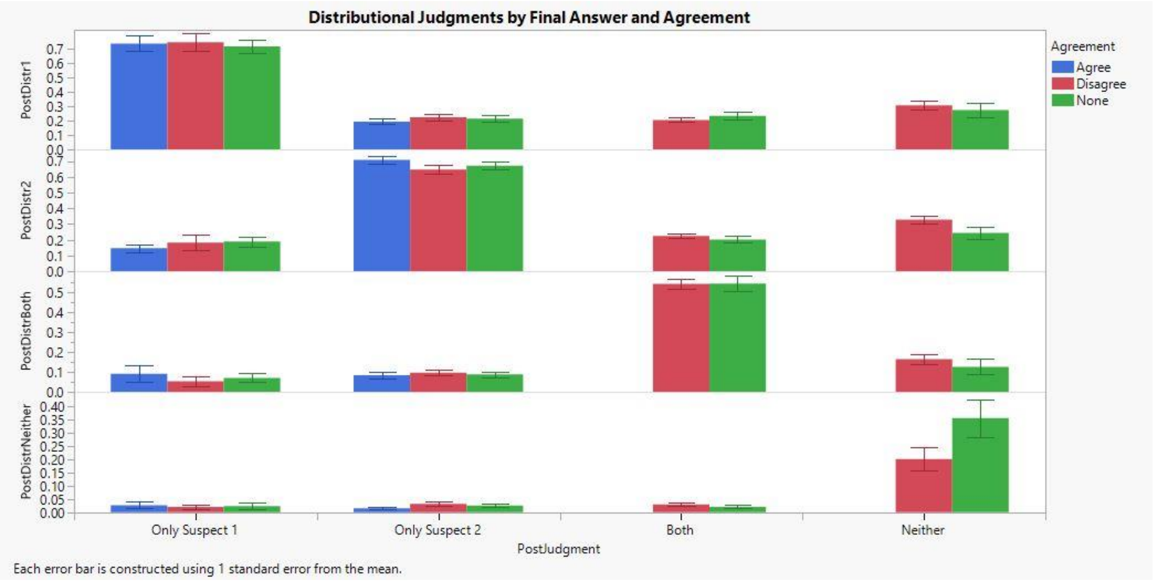


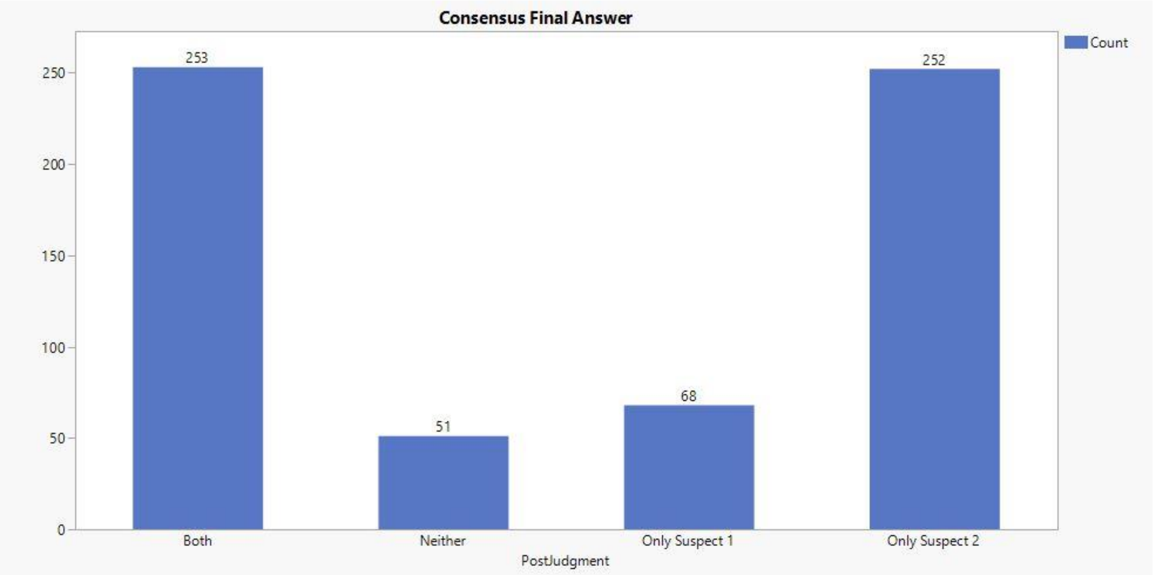
Figure 7. Predicted answer distributions by final answer.



Distributional Judgments by Final Answer and Agreement

Each error bar is constructed using 1 standard error from the mean.

Figure 8. Predicted answer distributions by final answer and agreement.



Consensus Final Answer

Figure 9. Distribution of final answers.

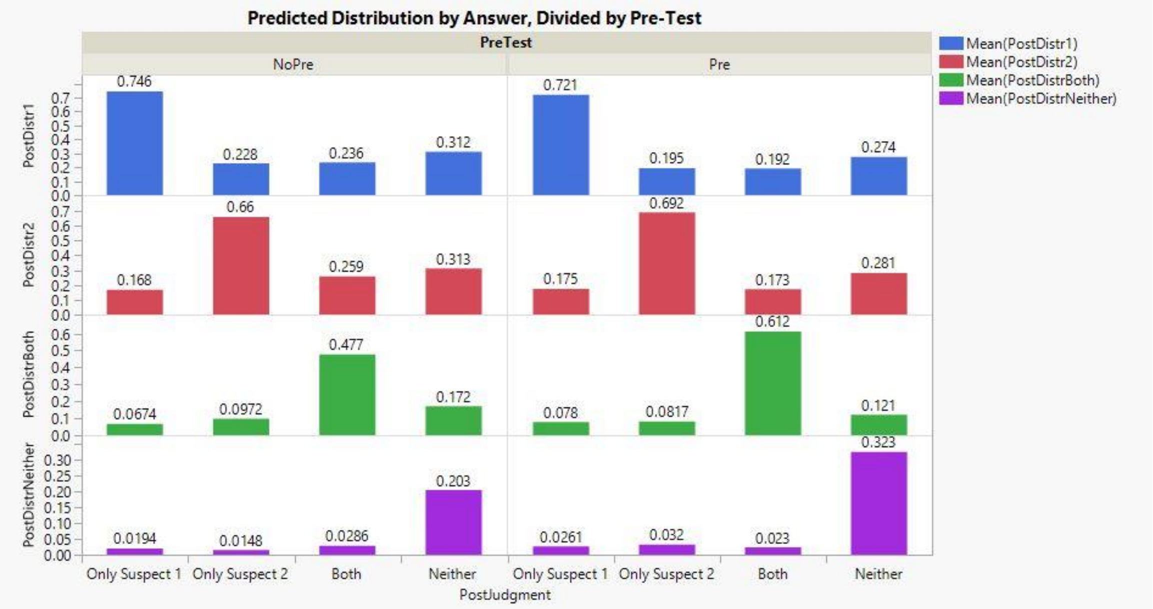


Figure 10. Predicted distribution by answer, divided by pre-test.

2.3 Discussion

We cannot reject the null hypothesis for H1, but the results are still exciting due to agreement level significantly affecting the truth serum as predicted. As mentioned prior, disagreement likely leads to a higher BTS score while agreement leads to a lower score. A large part of this observable effect results from participants' inherent bias towards their own answer preference. When estimating a population distribution—the second question in the Bayesian truth serum elicitation—participants are more likely to overrate the percentage of total participants that agree with them. Not only is this bias theorized in Prelec's 2004 paper, but it is also noticeable in this experiment (**Figure 7**). When participants disagree with their social influence, their distributional judgments change. **Figure 8** shows the difference in distributional judgments per answer depending on the agreement level. Participants who answered "neither" show the largest change in predicted distributions. I theorize that disagreement shifts the predicted distribution toward truth, whereas agreement shifts it away from the truth.

The shift in predicted distributions would explain the improved prediction score, as seen in **Figure 3**. I believe the prediction score is more predictive of disagreement's positive effect than the information score, as the levels of influence limit the agreement analysis in this particular design. Since we only used influence profiles that supported suspect one or suspect two, participants can only agree if they answered one of those two options. In this case, the average information score for the disagreement group is being pulled in a highly positive direction due to options "both" and "neither" having higher total BTS scores—including the information score—than options "only suspect one" and "only suspect two."

Figure 2 shows that participants who did not receive influence—and therefore could not agree or disagree—received scores on average that fell between the agree and disagree conditions. On the one hand, this plot supports the hypothesis that disagreement increases scores, as scores for participants who disagreed were far greater than scores for participants who received no influence. That said, the large divide between agree and disagree scores may be inflated by the participants with the best scores being unable to agree. Finally, **Figure 6** supports the disagreement hypothesis, with participants who answered "both" and "neither" receiving better scores if exposed to social influence, and consequently disagreed with the influence. This research design does not necessarily act as a limitation but seems to muddy some of the conclusions we can make. As a result, a future experiment including social influence profiles for each answerable option is imperative to further explore the effect of disagreement.

We fail to reject the null hypothesis for H2, as the statistical tests examining the pre-test condition were all significant. This finding is interesting, and after further investigation of the data, the test's significance can be attributed to participants anchoring on their pre-test. As seen on **Figure 10**, participants that receive a pre-test overestimate the distributions for answer options three (both) and four (neither). This overestimation is likely the result of anchoring on their pre-test instead of fully incorporating the information they receive from the influence in the middle of the experiment and adjusting their predictions accordingly. The result is surprising, and merits further examination into how pre-tests may affect Bayesian truth serum scores.

There is no significant difference in disagreement between participants who received low-quality data and those who received high-quality data, so we fail to reject the null for H3. However, there is an observable difference in agreement levels as seen on **Figure 5**. These findings suggest the need for further investigation; an experiment fully designed around information quality and social influence may find significant results.

The results of the main multinomial model found that influence significantly affects answer preference. Data quality and the influence by quality interaction were significant as well, indicating that depending on the data quality condition, influence has a larger effect on answer preference. While these results do not provide any support for H3, they suggest that evidence quality played a significant role in determining answer preference, especially when combined with different types of influence. This finding is further supported by the influence-only model, which was not statistically significant when evidence quality was not included.

We did not find any evidence of the truth serum's predictive ability in this experiment. The correct answer in the vignette is Suspect 2; as seen on **Figure 6**, respondents answering Suspect 2 received the third lowest truth serum scores. In this case, consensus outperforms the Bayesian truth serum, with the correct answer receiving the second most votes—off by just one vote for first—and 40% of the vote overall (**Figure 9**). In this task context, the Bayesian truth serum failed to detect the correct answer with the same efficacy as in the literature. Further experimentation would be beneficial, but currently our results cast doubt on the truth serum's effectiveness in solving crimes.

There are some possible limitations that could restrict the conclusions we can draw from this experiment. The existence of the "both" option complicates the agreement label for participants who chose that option. When receiving influence, participants who answer "both" are not necessarily agreeing with the influence, especially since some of the influence provides evidence against the alternative suspect. For this experiment, we coded answering "both" as disagreeing with the influence for that reason. However, there are surely participants that agreed with their influence and answered "both." To address this concern, we conducted an ANOVA predicting BTS score with agreement on data excluding any participants who answered "both." The results were significant, with an F-value of 9.265 ($p = 0.00011$). This statistical result supports the disagreement hypothesis and casts doubt on the idea that the agreement coding for participants who answered "both" had a significant confounding effect on this study's main findings. However, including "both" as an answerable option while not all influence profiles argue against the alternative answer is still a limitation when examining disagreement.

A final limitation would be the variance between the influence profiles. Because profiles were constructed using pilot data, there is no standard format that each influence profile follows. While this might reflect a real-world scenario where no social influence is the same, it does not necessarily help analyze the data and make generalizable conclusions. Some influence profiles may have provided more convincing arguments than others, possibly making it difficult to fully compare influence levels.

This experiment contributes to the literature in several ways. The recent Bayesian truth serum literature primarily focuses on testing the measure in a forecasting context, so this experiment provides some evidence for BTS's efficacy in predicting the correct answer in a different task. Currently, no experimentation has been conducted examining the BTS in a crime scene task. Furthermore, this experiment provides evidence for social influence affecting answer preference, building upon some of the results found in the social influence literature. That said, we did not find a significant effect of social influence on the truth serum; this experiment is currently the first to examine social influence's effect on BTS.

CHAPTER 3. EXPERIMENT TWO AND FOLLOW-UPS

3.1 Experiment 2a

The hypotheses for the second experiment are as follows:

• H4: The main effect of social influence will be significant; BTS scores will be lower in the experimental condition due to the decreased independence between participants injuring the surprisingly common signal.

• H5: Although participants in the experimental condition will have lower overall BTS scores, their prediction scores—the second half of the BTS calculation—will be higher, due to social influence causing predicted distributions to be more accurately calibrated.

This experiment builds upon the first experiment in a few ways. Along with the first experiment, it helps address the lack of novel experiments examining decision-making within the social influence literature. Likewise, it also seeks to contribute to the Bayesian truth serum literature by applying the truth serum to a psychology experiment with manipulations. This experiment involves a variety of tasks as well, all of which are less complex than the task involved in Experiment 1 but still seek to understand the same underlying processes. The three question formats—general knowledge, forecasting, and counterfactual—can all illuminate different ways and contexts that social influence affects decision-making. Additionally, as with Experiment 1, each question format investigates different processes used in intelligence analysis. Currently, experimentation relevant to techniques used in intelligence analysis is highly important, as most strategies employed by intelligence analysts are scientifically untested. Finally, the simplicity of this experiment—compared to Experiment 1—allows for us to make conclusions about social influence's impact on the different components of the truth serum.

A Monte Carlo simulation was conducted to support the plausibility of the two hypotheses. The simulation employed a question with only two options, A and B, and social influence consisting of an aggregate of 5 randomly sampled participants. Certain assumptions were built into the simulation, such as participant bias influencing their population distribution and random noise affecting predictions.

Two plots presenting the information and prediction scores as they relate to difficulty and social impact are included in the appendix. The difficulty parameter is simply defined by how many people out of 100 answered option A; this parameter mainly demonstrates how the BTS components move when most of the population prefers one option versus when it is split. The social impact parameter is how heavily a participant weights the social influence they receive when judging the population distribution; participants with 0 social impact do not use the influence at all, whereas participants with 1 social impact only follow the influence, and not their own individual decision-making. The two plots show that social influence injures the surprisingly common signal (information score) and improves the prediction score.

3.1.1 Methodology

3.1.1.1 Design and Participants

This experiment only varied whether participants receive social influence or not. Participants were recruited using the Qualtrics sampling system and rewarded with monetary compensation. In total, we gathered data from 316 participants.

3.1.1.2 Materials

This experiment was conducted using a simple Qualtrics survey. There were three questions about state capitals. Participants were asked if a city is a state's capital and could answer yes or no, and then rated their confidence in their answer. On each question, participants were asked what percent of people in the population will agree with them. Participants were also given a question in which they had to forecast the number of deaths related to COVID-19 by the end of December 2020. There was also one additional counterfactual forecasting question. Examples of each question type are located in the appendix.

In the social influence condition, participants received some information about how their peers answered in the following format: "5 previous participants from this study were randomly sampled. On the next question, 0 / 5 of them answered Yes.". In this experiment, participants were given a random sample of how five people answered. This information was sampled from pilot data and ecologically assigned, rather than randomly assigned. First, we ran a short pilot study identical to the control condition. Then, we created 100 random samples of five participants without replacement, per question. We calculated the frequency of "yes" answers per five-person sample, and then calculated the overall distribution of five-person samples. Finally, we manipulated the probability of the experiment assigning a participant to a sampling condition based on the observed distributions. For example, if 3/100 five-person sets on question one contained zero "yes" answers, participants had a three percent chance to receive a zero out of five in their influence on that question.

3.1.1.3 Procedure

There was only one phase of the experiment. After giving informed consent, participants completed the survey. It took no longer than 15 minutes.

3.1.2 Results

BTS scores were calculated for all five questions and then plotted. **Figure 11** shows the BTS scores across all questions and samples. Because the experimental condition assignments in this condition were probabilistic based on pilot data, some questions had zero participants in certain sampling conditions, as seen on the graph. As seen on the plot, there are possible inflection points in the counterfactual and forecasting questions, suggesting a linear trend, especially with the counterfactual. The effect size of BTS scores between experimental and control groups was 0.04, indicating a small effect.

Figure 12 shows the BTS scores divided among question and conditions. The results shown in **Figure 12** do not support H4, the hypothesis that BTS scores will be lower in the experimental condition. As seen in the plot, only two out of the five questions had lower BTS scores for the participants in the experimental condition.

Additionally, **Figure 13** plots the average prediction score of each condition per question. This plot does not support H5, as only two of the five questions yielded a better prediction score in the experimental condition. One interesting note is that the participants in the experimental condition performed better overall compared to their control counterparts on the Philadelphia question, but had less accurate distributional predictions. This indicates a stronger surprisingly popular signal than for the other questions, meaning that the social influence possibly could have heavily misinformed the experimental participants away from truth.

Figure 14 shows the BTS scores across all conditions and answers for the four verifiable questions. The number of questions is limited, but BTS performed badly, with the highest scores only predicting truth half of the time. BTS compared to consensus very poorly, which predicted the correct answer 100% of the time.

Two binomial regression models were created, one using all questions in the experiment and one only using the trivia questions. Both models predicted answer preference using sample and question. **Tables 9** and **10** show the likelihood-ratio test conducted for each of these models. Both tests were significant, demonstrating that both models perform significantly better than the null model.

This experiment provided promising insight into the changes that social influence may generate within the decision-making tasks at hand. However, due to the limited number of questions employed, it is difficult to make any substantial conclusions. Additionally, the ecological sampling design presents a unique limitation that unevenly assigned participants to each condition. The lack of evenly distributed participants across all sampling conditions and questions severely limited our ability to uncover social influence's potential effect. Therefore, three follow-up experiments were conducted, one for each question type.

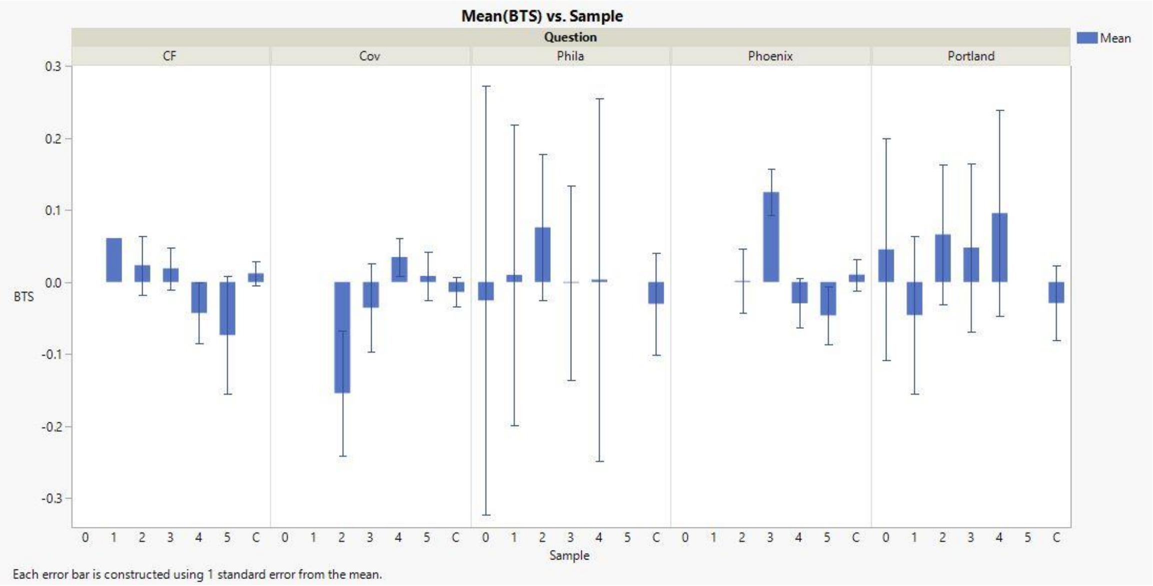


Figure 11. Plot of BTS scores across all samples and questions. CF represents the counterfactual question; Cov is the forecasting question; Phila, Phoenix, and Portland are the trivia questions.

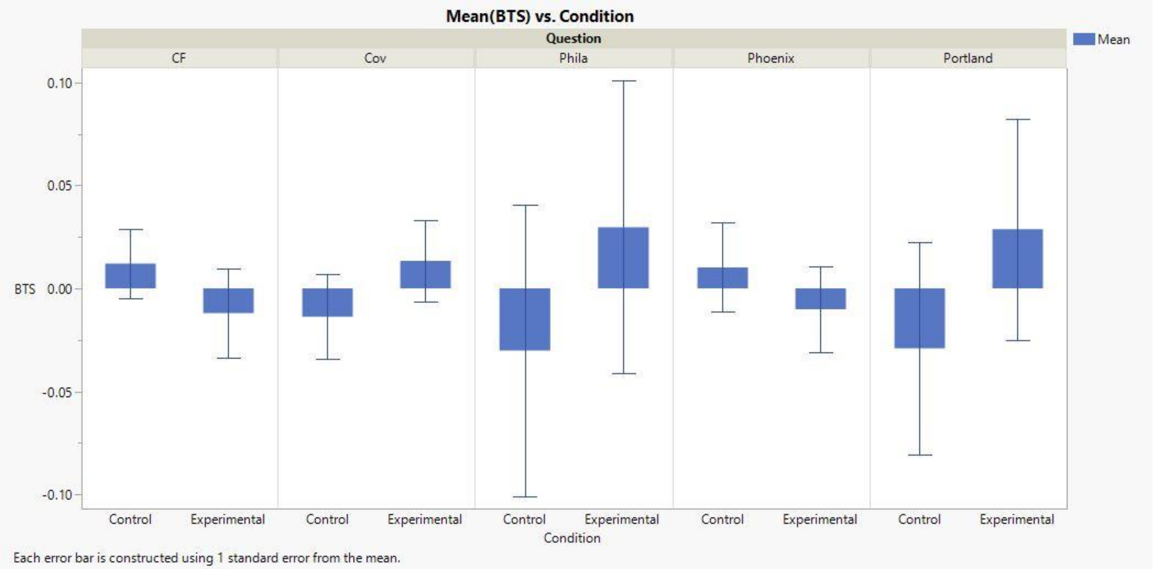


Figure 12. Plot of BTS scores broken up by condition per question.

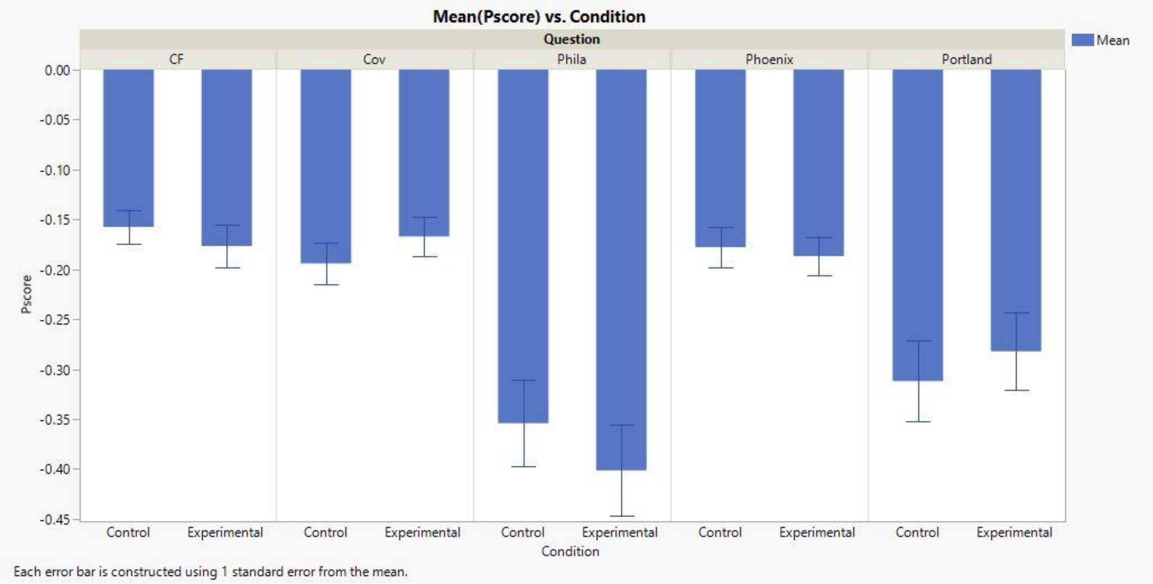


Figure 13. Plot of prediction scores broken up by condition per question.

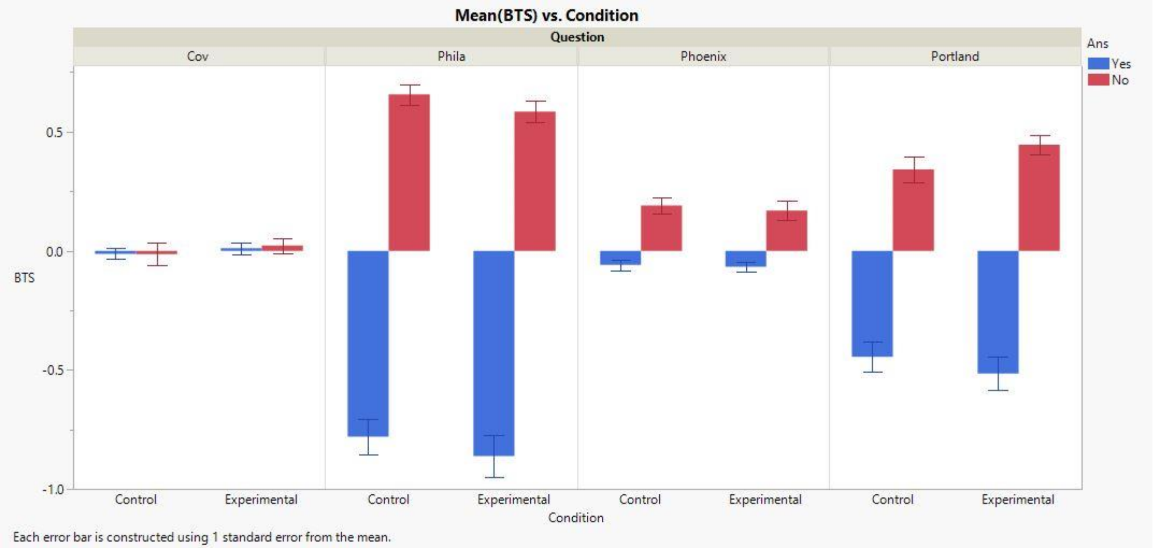


Figure 14. Plot of average BTS scores across all conditions and answers for all verifiable questions.

Table 9. Likelihood-ratio test for a binomial regression predicting answer preference using sample and question for all questions.

DF	*LogLik*	*ChiSq*	*P-value*
28	-970.4215	NA	NA
1	-1052.8644	164.8859	< 0.0001***

Table 10. Likelihood-ratio test for a binomial regression predicting answer preference using sample and question for only trivia questions.

DF	*LogLik*	*ChiSq*	*P-value*
17	-607.549	NA	NA
1	-653.891	92.68404	< 0.0001***

3.2 Experiment 2b

3.2.1 Methodology

3.2.1.1 Design

The goal of this experiment was to replicate and expand upon the trivia questions in experiment 2a. Like the previous one, this experiment varies whether participants receive social influence. Participants receive a social influence sample per question, which is randomized each time. The options for social influence range from 0 to 5; the influence is formatted as follows, using 0 as an example: "5 previous participants from this study were randomly sampled. On the next question, 0 / 5 of them answered Yes."

3.2.1.2 Participants

525 participants were recruited from the Qualtrics online participant pool. Each participant was given financial compensation determined by Qualtrics upon completion of the experiment. Participants were also instructed that the three best overall scores and three

best prediction scores would be given an additional reward of 15 dollars. 25 participants were excluded due to missing data for a final total of 500 participants.

3.2.1.3 Materials

The experiment was conducted using a Qualtrics survey. After opening the survey and giving informed consent, participants were randomly assigned to the control or experimental condition. Participants in the control condition answered ten state capital trivia questions. An example question is as follows: “Is Philadelphia the capital of Pennsylvania? Yes or No.” Along with each question, participants gave a confidence rating in their answer on a scale of 50-100, then answered how many participants they think will answer yes. Participants in the experimental condition answered the same set of questions, but before each question were given social influence. The social influence follows the same format as experiment 2a, but is randomized instead of being ecologically sampled. Participants in the experimental condition receive a different random sample for each of the ten state capital questions.

3.2.1.4 Procedure

After participants gave informed consent, they completed the survey. The responses took between 10 – 15 minutes.

3.2.2 Results

Figure 15 plots the BTS scores across both conditions for all ten questions. Of note, the control sample outperformed the experimental sample in every single question. Additionally, a linear mixed-effects model (**Table 11**) confirmed that the difference in BTS scores between the control and experimental conditions was significant. The effect size of BTS scores between experimental and control groups was -0.16, indicating a small negative effect.

Figure 16 shows the difference in prediction scores between conditions, with the experimental condition performing worse on all questions. The lower prediction scores provide a simple explanation for why BTS scores were lower in the experimental condition: participants were not as accurate in their predicted sample distributions when receiving social influence. In this non-ecological, randomly assigned experiment, social influence disinformed participants instead of helping to calibrate their predictions properly.

A second version of the BTS was calculated using the control group as a reference class. In the control-referenced version, the surprisingly popular score was calculated only using the control group and then assigned to participants in the experimental conditions depending on their answers. The prediction score was calculated by comparing the experimental participants' predicted distributions to the control group's actual distribution. This BTS variant was developed to discover how misinformation, disinformation, or social influence may affect the score.

A plot of the control-referenced BTS variant demonstrates the differences social influence and possible misinformation can create between conditions (**Figure 17**). In this plot, the control group's mean BTS stayed close to zero while the experimental conditions showed considerable differences. I hypothesize that this is due to the prediction score reflecting inaccuracies in the experimental conditions' distributional judgments—since social influence has altered the ability to judge how a population will answer accurately.

Table 12 presents a linear mixed-effects model that predicted BTS scores with the experimental condition and designated subject as a random effect. The results of that model were statistically significant, with a χ2 value of 17.2983 and a p-value of .0082, indicating that participants in different sample conditions had significantly different BTS scores, regardless of the question. A binomial regression predicting answer preference using sample and question yielded insignificant results, aside from the main effect of question.

A linear mixed-effects model was conducted that predicted Brier scores with confidence and question, and designated subject as a random effect. As seen in **Table 13**, the results were significant across the main effect and the interaction, suggesting that confidence is predictive of answer accuracy, depending on the question.

As with other studies examining the surprisingly popular signal's effectiveness, correctness was compared across different aggregation methods. **Figure 18** plots the performance of BTS (SP, or surprisingly popular), consensus, and confidence-weighted consensus. With the SP method, the highest average BTS score was taken between each of the binary choice options for each question. Consensus chose the most popular answer among participants per question; confidence-weighted consensus multiplied the average confidence rating per answer by the percent of participants who submitted that answer. In our ten question sample, the BTS performs worse in almost all sampling conditions. Future replication that employs all fifty state capitol questions is recommended to further elucidate the damaging effect of social influence.

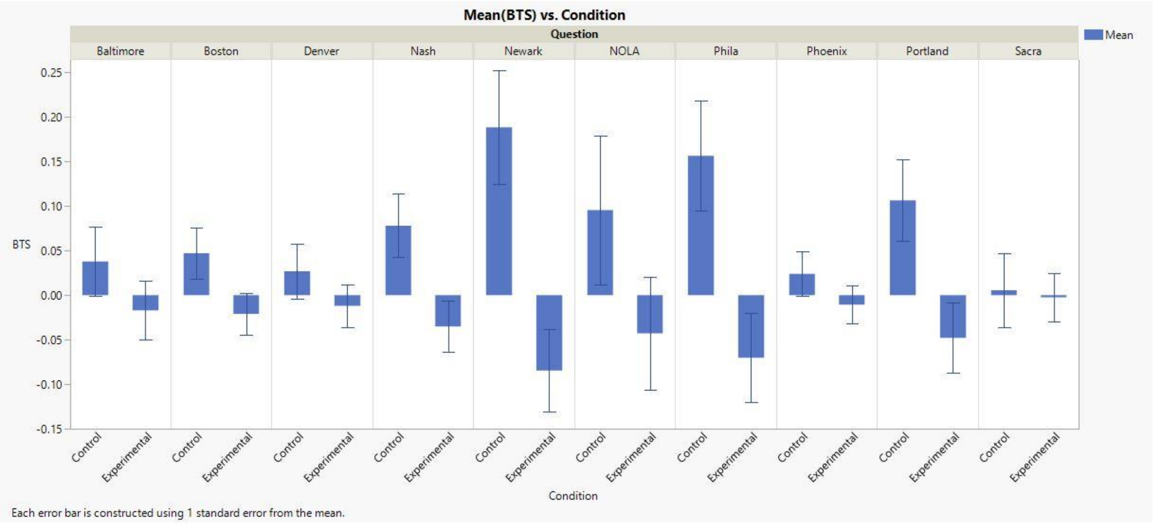


Figure 15. Plot of BTS scores across all questions and conditions.

Table 11. Linear mixed-effects model predicting BTS scores with condition and question, designating subject as a random effect.

	Chi Sq	*DF*	*P-value*
Intercept	0.4814	1	0.487776
Condition	0.6977	1	0.403550
Question	16.0549	9	0.065745
Interaction	23.2679	9	0.005622***

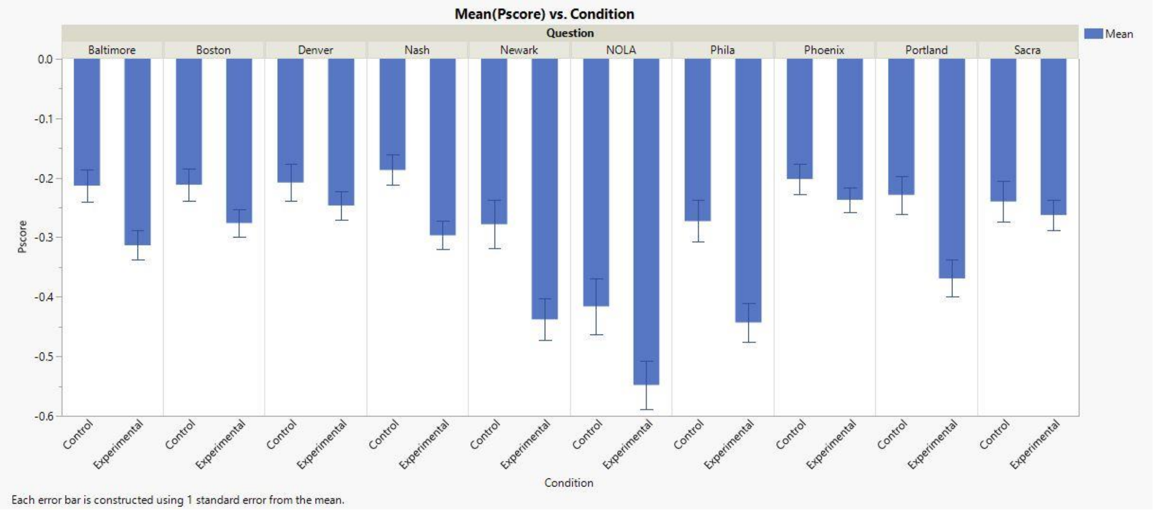
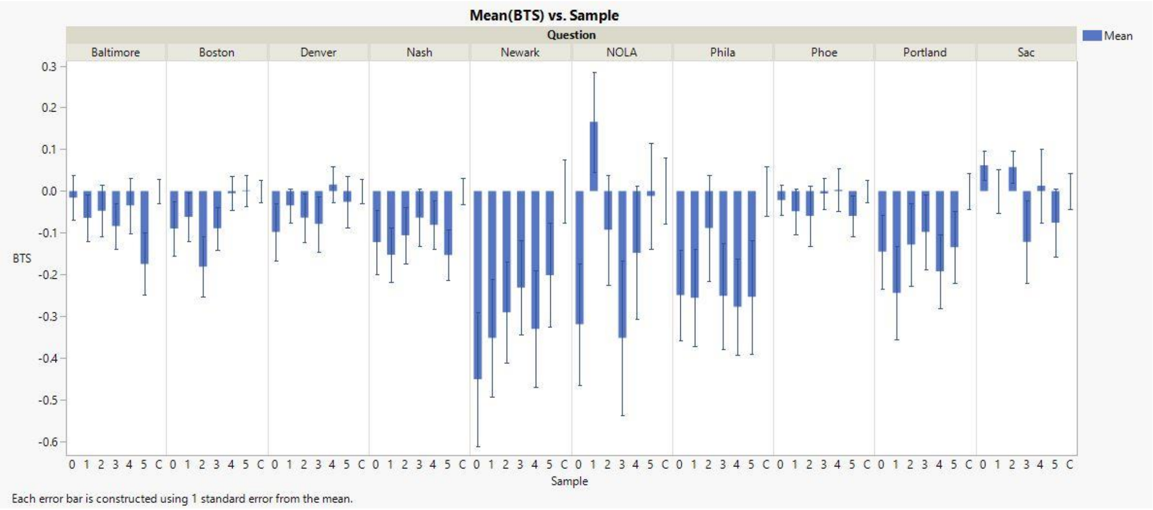


Figure 16. Plot of prediction scores across all questions and conditions.



Each error bar is constructed using 1 standard error from the mean.

Figure 17. Graphing control-referenced Bayesian truth serum variant vs question and sample. Sampling condition C is the control.

Table 12. Results of fitting a linear mixed-effects model predicting BTS with Sample, designating Subject as a random variable.

	ChiSq	*DF*	*P-value*
(Intercept)	6.92122	1	0.0085179**
Sample	17.29827	6	0.0082474**

Table 13. Linear mixed-effects model predicting Brier scores using confidence and question, designating subject as a random effect. Analysis of Deviance table.

	ChiSq	*DF*	*P-value*
(Intercept)	0.0119	1	0.913
Confidence	31.2979	1	< 0.0001 ***
Question	39.5135	9	< 0.0001 ***
Conf*Question	95.6793	9	< 0.0001 ***

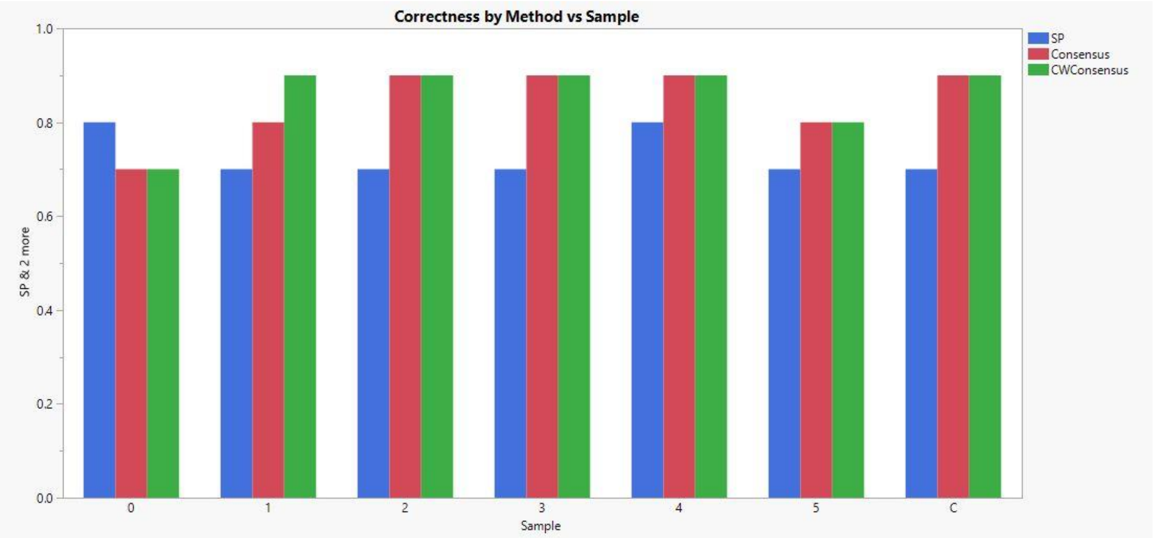


Figure 18. Plotting percent correct (out of ten questions) by aggregation method.

3.3 Experiment 2c

3.3.1 Methodology

3.3.1.1 <u>Design</u>

The goal of this experiment was to replicate and expand upon the singular forecasting question asked in experiment 2a. This experiment only varies whether participants receive social influence or not. The social influence sample that participants receive is randomized per question.

3.3.1.2 <u>Participants</u>

526 participants were recruited from the Qualtrics online participant pool. Financial compensation, including the bonus for high performance, was the same as in experiment 2b.

3.3.1.3 Materials

This experiment was conducted using a Qualtrics survey. This time, participants answered forecasting questions from various topics. Examples include: “By the end of December 2020, will there be more than 300,000 deaths in the United States as a result of COVID-19, up from 143,000 as of July 2020?” and “Will Jacinda Ardern win the 2020 New Zealand General Election?” Like the other experiments, participants also give a confidence rating and a prediction of how many other participants will answer yes. Participants were divided based into control or experimental conditions. Participants in the experimental condition received social influence in an identical format to experiment 2b.

3.3.1.4 Procedure

After participants gave informed consent, they completed the survey. The survey responses took 10 – 15 minutes.

3.3.2 Results

Similar analyses to experiment 2b were conducted with the data from this experiment. A linear mixed-effects model that predicted BTS with question and condition found insignificant main effects and interactions (**Table 14**). The effect size of BTS scores between groups was -0.03, indicating a small negative effect. A linear mixed-effects model that predicted BTS score with sample and designated subject as a random effect was found significant (**Table 15**). The model’s $\chi 2$ value was 21.403, with a p value of .0015. The results from this model suggest that participants had significantly different BTS scores in different sample conditions, without controlling for question. A binomial regression predicting answer preference using sample and question was found to be insignificant, aside from the main effect of question.

Figure 19 plots the BTS scores across all questions and conditions. The plot is promising despite the insignificant results, as five out of the seven questions show higher BTS scores in the control condition. These results are consistent with experiment 2b, as the control outperformed the experimental condition. While trivia and forecasting are different tasks involving different cognitive processes, the data suggests that BTS scores were injured by social influence in both experiments.

Like experiment 2b, a control-referenced BTS score variant was calculated for the forecasting questions (**Figure 20**). While more conservative than the plot shown in experiment 2b, this graph shows a similar trend of experimental conditions mostly scoring in the negative. The two conditions that showed better performances in **Figure 19** also appear positive in **Figure 20**, although some samples perform better than others.

Table 14 shows the results of a linear mixed-effects model predicting Brier scores using confidence and question. The interaction was significant, indicating that confidence level significantly predicted answer accuracy, depending on the question.

Figure 21 shows overall correctness by method for all sampling groups, including the control. Similarly to experiment 2b, taking the answer with the highest average BTS score per question yielded less correct answers than simply taking the consensus. Noticeably, the control condition performed just as badly as the experimental conditions, ruling out the possibility that the lower accuracy is associated with social influence in this case. The literature suggests that in forecasting contexts, either an extension is necessary to improve BTS for forecasting, or the sample should only include people with high domain knowledge on the forecasting topic to improve accuracy (Lee, Danileiko, & Vi, 2018; Olsson, de Bruin, Galesic, & Prelec, 2019; Rutchick, Ross, Calvillo, & Mesick, 2020). This experiment included neither, adding to the already existing evidence that BTS does not always perform accurately in forecasting contexts.

Table 14. Linear mixed-effects model predicting BTS scores using condition and question, designating subject as a random effect.

	Chi Sq	*DF*	*P-value*
Intercept	0.2837	1	0.5943
Condition	0.4034	1	0.5254
Question	3.4547	6	0.7500
Interaction	4.9113	6	0.5552

Table 15. Linear mixed-effects model predicting BTS scores using sample, designating subject as a random effect.

	ChiSq	*DF*	*P-value*
(Intercept)	7.139392	1	0.0075409
Sample	21.403358	6	0.0015522***

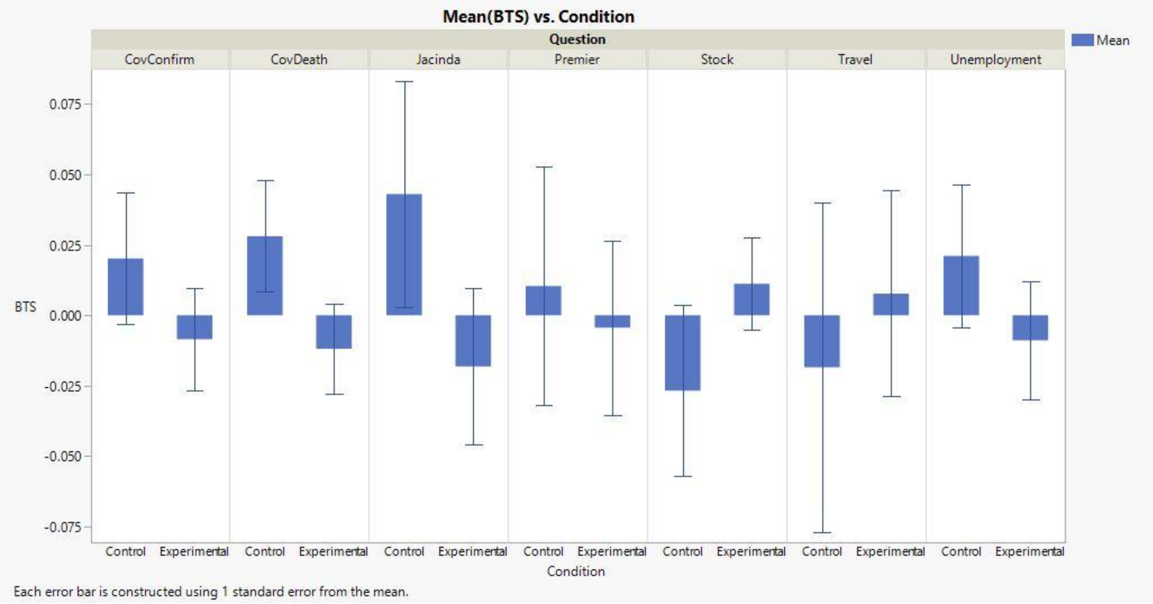


Figure 19. BTS scores across all questions and conditions.

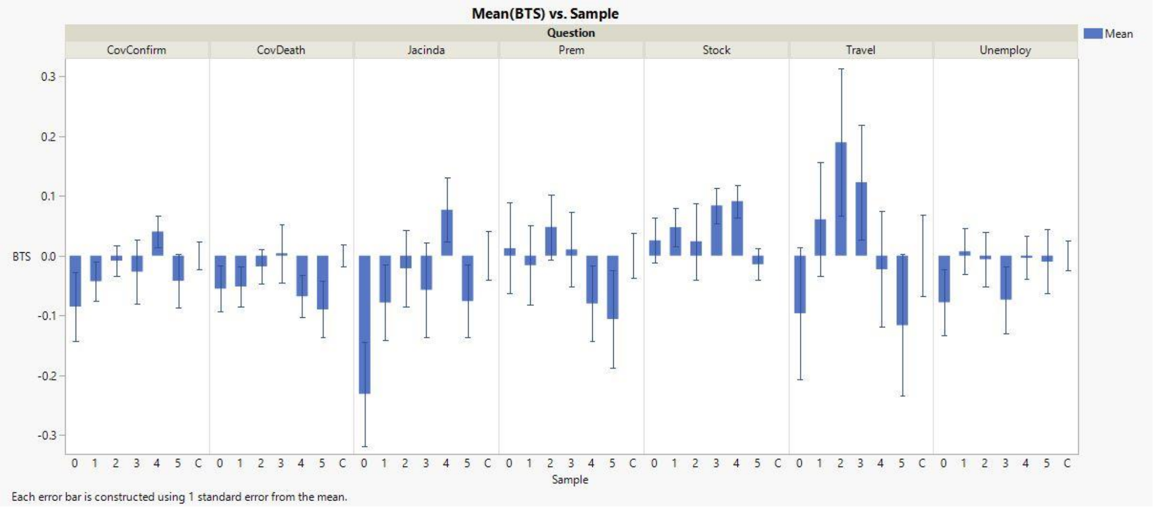


Figure 20. Control-referenced BTS variant plotted across all questions and samples.

Table 16. Linear mixed-effects model predicting Brier scores using confidence and question, designating subject as a random effect. Analysis of Deviance table.

	ChiSq	*DF*	*P-value*
(Intercept)	18.8381	1	< 0.0001***
Confidence	3.6103	1	0.05742
Question	62.8890	6	< 0.0001***
Conf*Question	125.8875	6	< 0.0001***

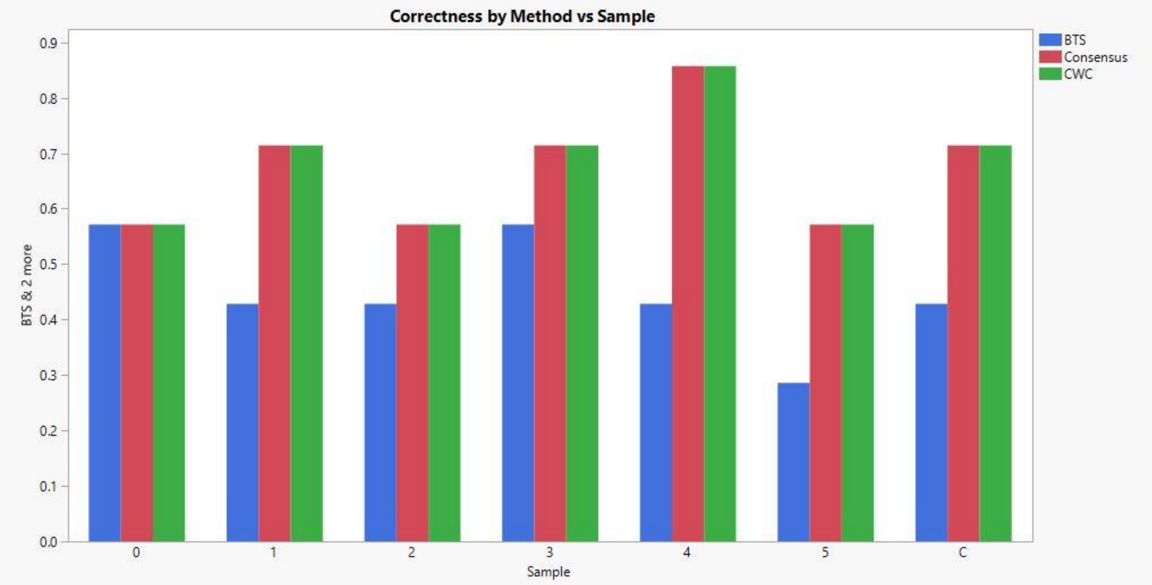


Figure 21. Plotting percent correct (out of seven questions) by aggregation method.

3.4 Experiment 2d

3.4.1 Methodology

3.4.1.1 <u>Design</u>

The goal of this experiment was to replicate and expand upon the counterfactual question asked in experiment 2a. As with the others, this experiment varies whether participants receive social influence or not. The social influence that participants receive is randomized.

3.4.1.2 Participants

525 participants were recruited from the Qualtrics online participant pool. Financial compensation, including the possible bonus reward, were the same as the previous experiments.

3.4.1.3 Materials

This experiment was conducted using a Qualtrics survey. Participants answered three counterfactual questions about historical events. These questions were longer with more uncertainty than the questions in experiments 2b and 2c. An example counterfactual question is included in the appendix. The structure and format of the questions participants answered were identical to the other experiments. The social influence conditions were also assigned randomly again as well.

3.4.2 Results

Again, similar analyses to the previous experiments were conducted with the counterfactual data. A linear mixed-effects model predicting BTS score with condition and question yielded insignificant results (**Table 17**). The effect size was -0.12, indicating a small negative effect. Additionally, **Figure 22** plots the BTS scores per question and condition. As seen on the plot, participants in the experimental condition received lower scores than those in the control condition, on average. These results are in line with the previous experiments, showing that the possible disinformation from social influence negatively affected BTS scores—and thus, judgment ability.

As with the other experiments, a control-referenced BTS variant was calculated and plotted (**Figure 23**). The plot shows a similar trend of mostly negative experimental scores, supporting the initial results. The Donner question in particular shows a possible inflection point at influence group 3, where scores become increasingly negative as the social influence increases in the amount of people saying "yes." This is likely due to the actual distribution residing between sample groups 2 and 3, as 49% of participants answered yes on that question. Therefore, participants who received social influence not from groups 2 and 3 would have been heavily disinformed and vulnerable to over or underestimating the population's preferences. The same trends are seen on the plot of standard BTS scores, broken up by sample (**Figure 24**).

A binomial regression predicting answer preference with sample and question yielded significant results (**Table 18**). The main effect of sample was significant, indicating that answer preferences changed depending on the social influence. The main effect of question was also significant, indicating that answer preferences change between questions. Finally, the interaction was not statistically significant, meaning that answer preferences did not significantly change per sample, depending on the question.

Table 17. ANOVA predicting BTS score with condition.

	Chi Sq	*DF*	*P-value*
Intercept	1.7677	1	0.1837
Condition	2.5082	1	0.1133
Question	0.2131	2	0.8989
Interaction	0.3023	2	0.8597

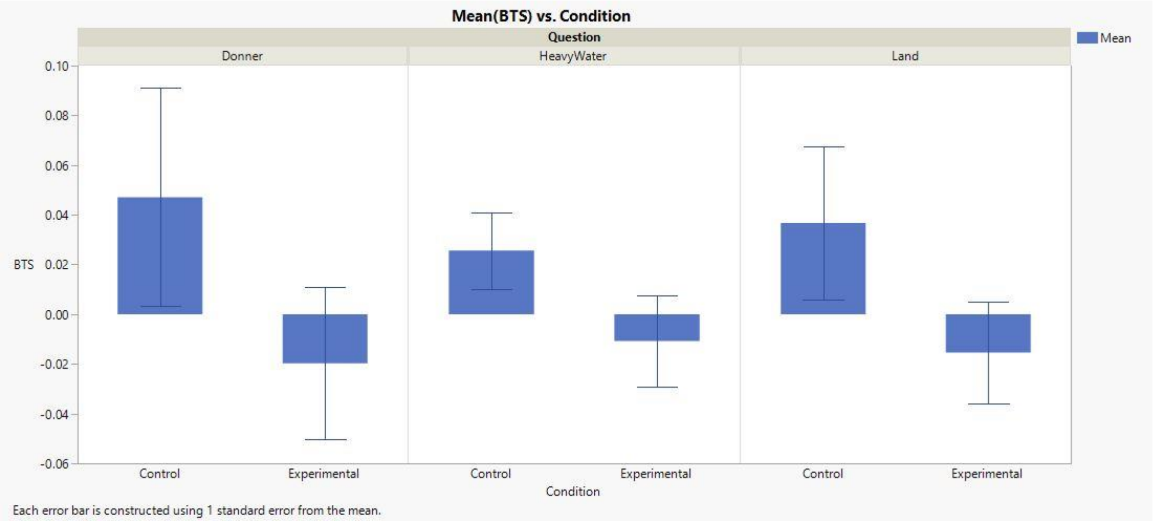


Figure 22. Plotting BTS score by question and condition.

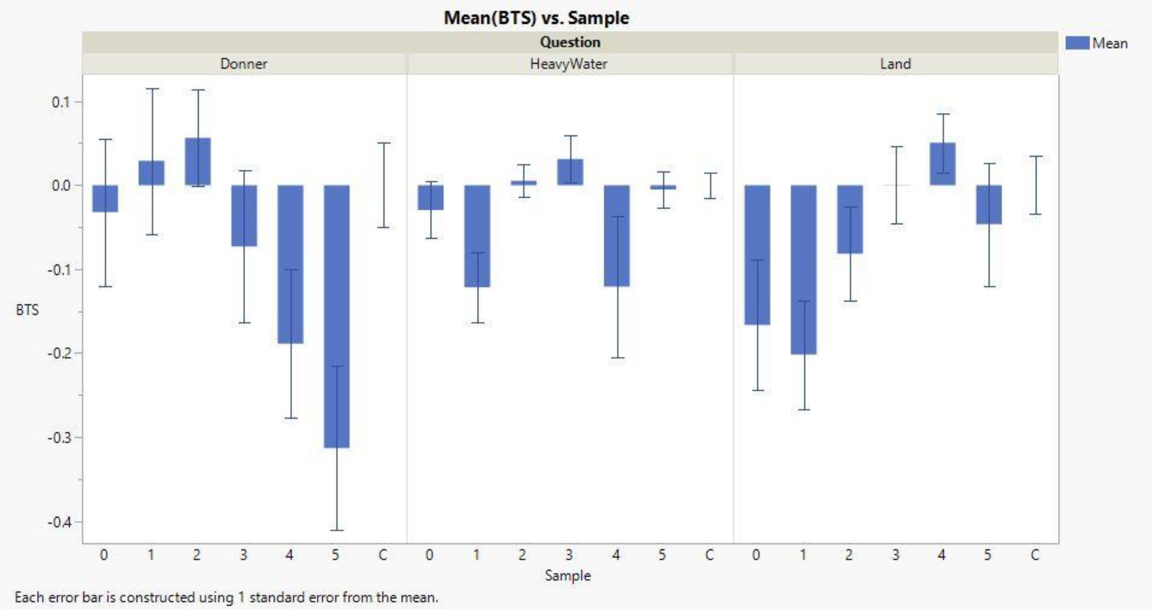


Figure 23. Control-referenced BTS scores.

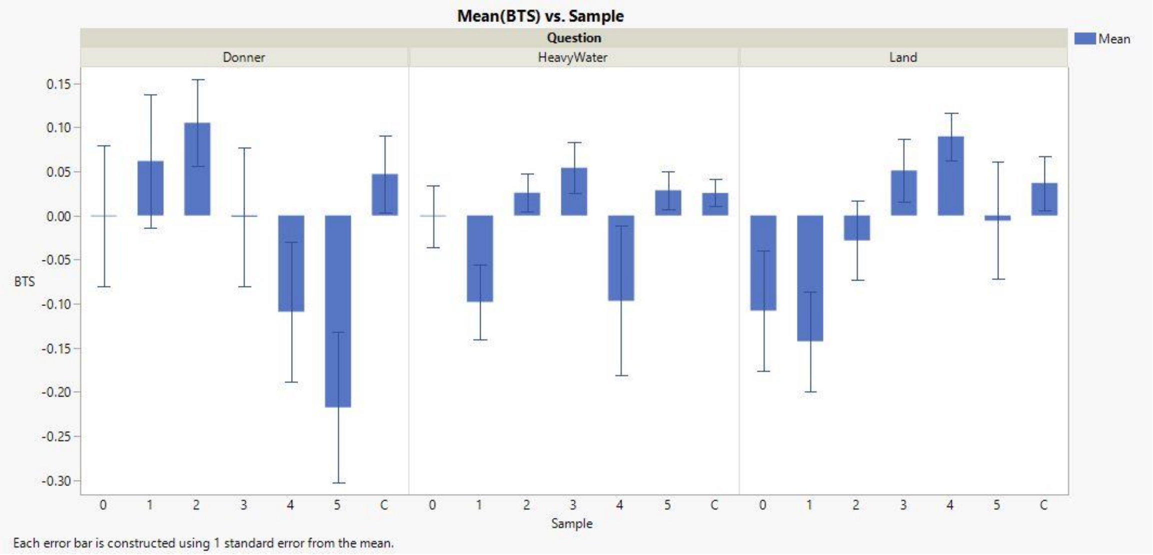


Figure 24. Standard BTS scores broken up by sample and question.

Table 18. Binomial regression predicting answer preference with sample and question; Likelihood-ratio test.

DF	*LogLik*	*ChiSq*	*P-value*
21	-798.48	NA	NA
1	-932.96	268.95	< 0.0001***

3.5 Experiment 2a-2d Discussion

The four experiments together provide promising results for elucidating the effect of social influence on the Bayesian truth serum. When answering general knowledge questions, social influence appeared to have a highly significant effect, with participants in the control condition receiving higher BTS scores on every single question. Whether social influence disrupts the predictive ability of BTS in this context is still unclear, as the participants in the control condition performed just as badly as the experimental participants (**Figure 18**).

While the main effect test in experiment 2c was not significant, **Figure 19** shows very promising results in that the control condition received better BTS scores on average than the experimental condition on five out of the seven questions. The insignificant findings may be due to many variables, as forecasting questions—particularly questions about a pandemic many laypeople do not understand—can be difficult without knowledge in the question's domain. Nonetheless, this finding builds upon the trend that the social influence implemented in these experiments is injurious to the truth serum. Once again, while the control condition had higher BTS scores on most questions, it performed badly—sometimes worse than some of the sampling conditions—on predictive accuracy.

The final experiment builds upon the prior experiments; while experiment 2d has insignificant results, **Figure 22** fits the trend of the other experiments. It is possible that an expanded version of experiment 2d with more counterfactual questions would yield significant results, as only three questions failed to reach significance—despite showing observable differences in the bar plot.

Although our hypotheses from experiment 2 did not necessarily translate to the follow-up experiments due to differences in experimental design, their results support H4. I theorize that in this case, social influence was injurious to the truth serum on average due to the influence largely being dis-informative over the aggregate. While experiment 2 probabilistically assigned participants a sampling condition based on our ecology found in the pilot data, experiments 2b-d randomly assigned participants to sampling conditions, drastically increasing the number of participants that would be exposed to social influence that was dis-informative. Additionally, looking at the control-referenced truth serum plot for experiment 2b, the experimental group receiving influence closest to the actual distribution performed better than random, receiving, on average, a rank of 2.3 out of 6 over all ten questions and the first-place rank on three of the questions. This finding further supports the theory that dis- or misinformation impairs the truth serum.

The follow-up experiments elucidated the possibility that BTS can detect disinformation or misinformation but showed little support for the truth serum's predictive ability—not just in the experimental condition, but, more importantly, in the control condition. Over fifty state capital questions, Prelec showed that the truth serum outperformed all other methods, including consensus and confidence-weighted consensus (Prelec, Seung, & McCoy, 2017). In our ten-question sample, the truth serum only performs better in one sampling condition. Perhaps over fifty questions, we would find a similar effect.

However, there are considerable methodological differences between experiment 2b and Prelec’s series of experiments: Prelec’s three state capital studies only recruited students from elite universities (two of the studies recruited MIT students, one study recruited Princeton students), whereas our participants were recruited randomly nationwide and represent the general population more accurately; across three experiments, Prelec’s team only recruited 116 participants, whereas we collected data from 500 participants. Additionally, for better or for worse, Prelec’s studies were conducted in person and sometimes with pen and paper. In contrast, our experiment was conducted entirely online—and not in a lab environment. A final small difference is that in Prelec’s state capital experiments, they always asked if the most populous city in a particular state was the capital. Our experiment is consistent with that methodology on all but one question, which asks if Sacramento is the capital of California. With those differences in mind and the large discrepancy in question number, it is difficult to make conclusions about social influence’s effect on the predictive nature of BTS compared to Prelec’s work. We asked Prelec for his raw data to compare to ours in light of these differences but received no response.

These experiments had several limitations which also impede our ability to make generalizable conclusions. First and foremost, the online format may lead to slightly lower quality data. At the same time, we implemented attention checks and thoroughly sifted through data to replace participants that did not pay attention; it is unclear if participants approach online studies with the same rigor as in-lab experiments. Additionally, a participant’s geographic identity may have had a significant effect on their trivia question answers. For example, 87.2% of participants correctly identified Sacramento as the capital of California, a much larger margin than any other question. Furthermore, the BTS forecasting-related literature theorizes that domain knowledge may be an important factor in strengthening the truth serum’s predictive capabilities (Lee, Danileiko, & Vi, 2018; Olsson, de Bruin, Galesic, & Prelec, 2019; Rutchick, Ross, Calvillo, & Mesick, 2020). It is likely that not all participants held expertise in every single domain featured in Experiment 2c, some of which are frequently unfamiliar to Americans, such as the New Zealand General Election and the English Premier League.

CHAPTER 4. GENERAL DISCUSSION

The Bayesian truth serum remains a measure with great potential, and further exploration into social influence's impact on the serum is undoubtedly necessary. Although some results were not statistically significant, there was pervasive evidence that social influence plays a disruptive role in decision-making and the serum over several task contexts. The results from these experiments suggest that social influence does not have a significantly injurious effect on the BTS's predictive ability, particularly because the control and experimental conditions both performed badly. Contrary to the literature, particularly Prelec, Seung, & McCoy (2017) where the BTS outperformed all other methods consistently, the BTS failed to accurately forecast in all but one sampling condition. These results are notable due to the control groups performing just as badly as the experimental groups. Expanded replications and future experimentation—specifically forecasting experiments that measure and incorporate domain knowledge—may elucidate BTS's capabilities in these task contexts more definitively.

Additionally, the significant main effect of pre-testing in Experiment 1 was exciting and surprising, and could lead to a novel branch of research on the BTS. Currently, no studies have investigated the possible effects of asking the BTS twice for one task, and our findings show promising results that a pre-test BTS may be damaging to the overall scores. Even when only comparing the post-test BTS scores between pre-test conditions, the post-test scores were anchored on the pre-influence BTS. I plan on conducting future research to further examine how pre-test BTS elicitations may damage Bayesian truth serum scores.

Although there were some limitations in each experiment, they provide inspiration for future experimentation. The small differences between influence profiles in Experiment 1 made it difficult to determine if one profile may be more convincing than the others. Regardless, Experiment 1 yielded promising results, providing evidence that disagreements may improve reasoning and BTS scores. Thus, future experimentation also testing the effects of disagreement is necessary, albeit with a different task and more standardized influence.

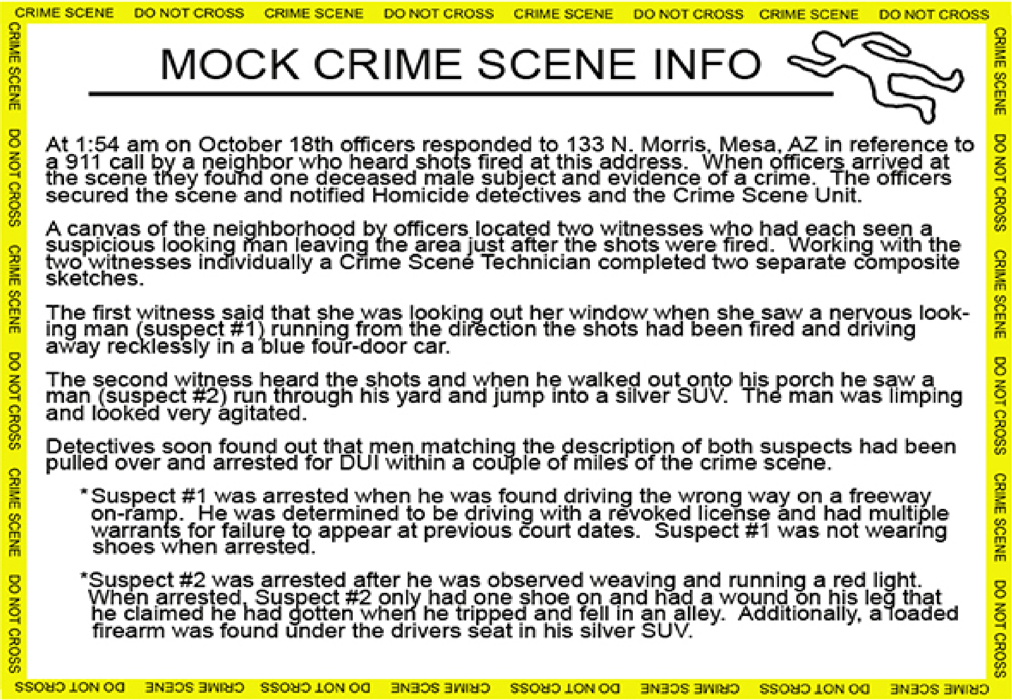
Experiments 2a through 2d were simpler and contained less limitations. However, future experimentation is recommended due to the differences in methodology, sample size, and final results, when comparing 2b to Prelec, Seung, & McCoy (2017)—if only to further test the discrepancy between this study and the existing literature. A future study utilizing an expanded, in-person state capital survey with a large amount of Georgia Tech students would be a more direct replication. Such an experiment could test the replicability of Prelec et al. (2017) with a different, larger participant pool, while also continuing to evaluate social influence's effects.

Perhaps the most compelling findings in this series of experiments are the truth serum's potential to expose differences between groups—such as disinformation—and disagreement's ability to improve reasoning about the population. While it requires additional testing, calculating a BTS variant that theoretically compares a disinformed sample to a control sample could have significant applications as technology continues to advance. Disinformation is just one example of this methodology's application; it could be used to compare two groups that differ on any one meaningful variable.

Furthermore, inducing disagreement to enhance the reasoning process could be critical in receiving better results in intelligence analysis and corporate decision-making. Sniezek & Henry (1989) found evidence of disagreement improving numerical judgment accuracy in a group-work setting. These studies build upon that finding in a social influence context—albeit with a different task. Not only do the effects of disagreement observed here yield a beneficial application across several domains to improve decision-making, but they also attest to the value of interacting with a set of diverse opinions. Interacting with and listening to opposite opinions and experiences, while difficult sometimes, adjusts our beliefs closer toward reality—as it did in Experiment 1.

The Bayesian truth serum is still young and relatively unstudied, but our series of experiments contribute some surprising and interesting results. Additionally, these experiments contribute new ideas to the social influence literature and introduce social influence into the BTS literature. This research program is fresh and exciting and can make long-standing contributions in the forecasting and decision-making domains.

APPENDIX A. EXPERIMENT ONE VIGNETTE


CRIME SCENE
DO NOT CROSS

MOCK CRIME SCENE INFO

At 1:54 am on October 18th officers responded to 133 N. Morris, Mesa, AZ in reference to a 911 call by a neighbor who heard shots fired at this address. When officers arrived at the scene they found one deceased male subject and evidence of a crime. The officers secured the scene and notified Homicide detectives and the Crime Scene Unit.

A canvas of the neighborhood by officers located two witnesses who had each seen a suspicious looking man leaving the area just after the shots were fired. Working with the two witnesses individually a Crime Scene Technician completed two separate composite sketches.

The first witness said that she was looking out her window when she saw a nervous looking man (suspect #1) running from the direction the shots had been fired and driving away recklessly in a blue four-door car.

The second witness heard the shots and when he walked out onto his porch he saw a man (suspect #2) run through his yard and jump into a silver SUV. The man was limping and looked very agitated.

Detectives soon found out that men matching the description of both suspects had been pulled over and arrested for DUI within a couple of miles of the crime scene.

*Suspect #1 was arrested when he was found driving the wrong way on a freeway on-ramp. He was determined to be driving with a revoked license and had multiple warrants for failure to appear at previous court dates. Suspect #1 was not wearing shoes when arrested.

*Suspect #2 was arrested after he was observed weaving and running a red light. When arrested, Suspect #2 only had one shoe on and had a wound on his leg that he claimed he had gotten when he tripped and fell in an alley. Additionally, a loaded firearm was found under the drivers seat in his silver SUV.

The Crime Scene Unit photographed the scene, collected evidence, and processed for latent prints. Overall, midrange, and close up photographs were taken of the items of evidence marked with placards. Each item with a numbered placard was collected as evidence and each lettered item was swabbed for DNA. All items were packaged and submitted to evidence for impound. The scene was then processed for latent fingerprints using Black and Magna Powders. Latent fingerprints were detected, lifted and submitted to the Latent Print Unit for further analysis and comparison to suspects #1 and #2.

APPENDIX B. EXPERIMENT ONE LOW QUALITY STIMULI

Blood and urine sample were collected from the two suspects. The Toxicology Unit analyzed the blood for alcohol using Headspace-Gas Chromatography (GC) with Flame Ionization Detection (FID), which resulted in a 0.12% blood alcohol concentration on suspect #1 and a 0.18% blood alcohol concentration on suspect #2.

The Urine was screened for drugs using Enzyme-Linked Immunosorbent Assay(ELISA.) Methamphetamine and amphetamine (methamphetamine metabolite) were confirmed by Gas Chromatography/Mass Spectrometry (GC/MS) in suspect #1. Suspect #2 had negative results.

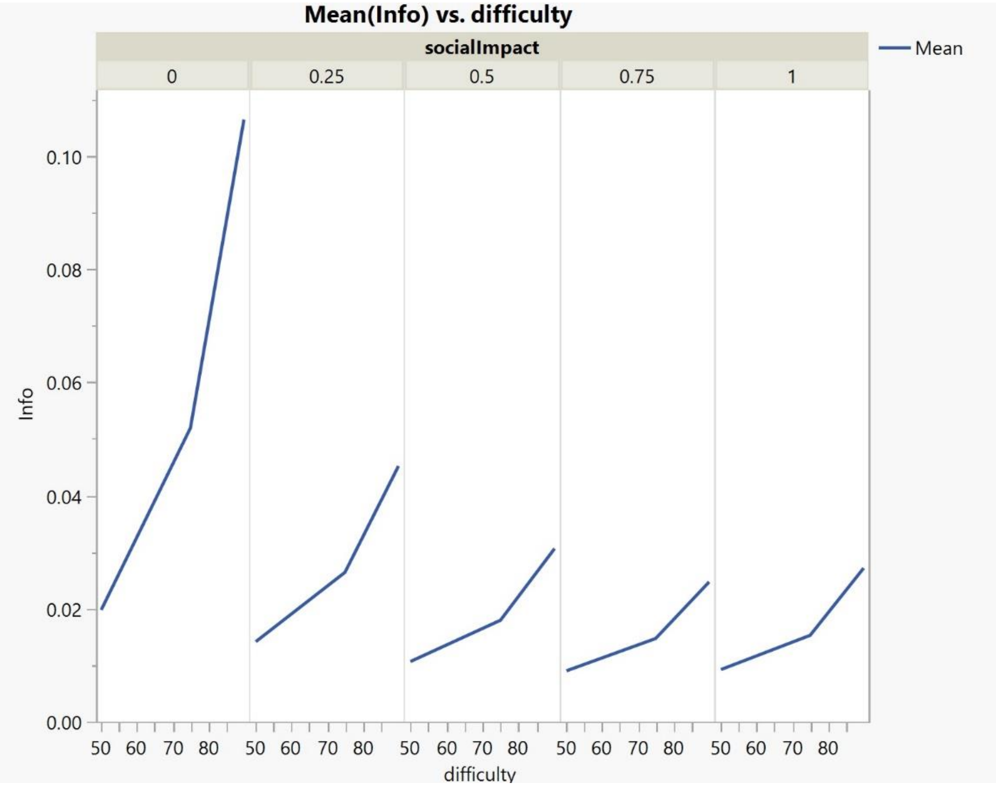
APPENDIX C. EXPERIMENT ONE HIGH QUALITY STIMULI

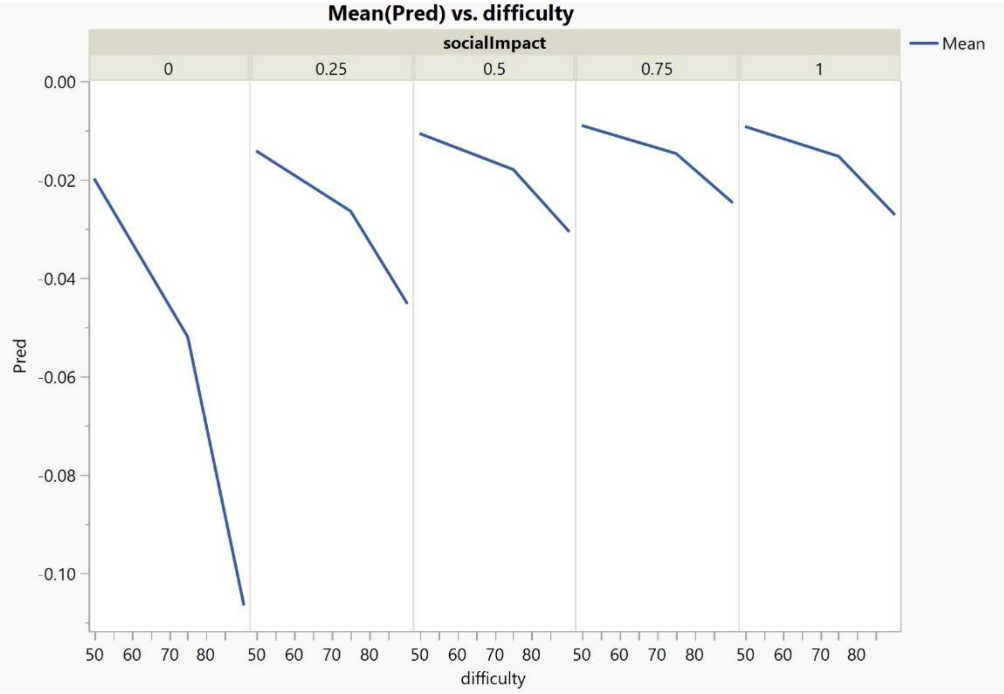
The Firearms Unit analyzed two cartridge cases, one bullet (projectile), and one firearm recovered from the crime scene. They also analyzed another firearm recovered from suspect #2. The Firearms Unit compared the bullet and cartridge cases from the crime scene to test fires from the firearm collected at the scene (item marked #6). It was excluded as having fired the bullet/cartridge cases. A serial number restoration was also completed on this firearm. Running the serial number through NCIC revealed that the weapon was stolen six months earlier. The Firearms Unit compared the bullet and the cartridge cases from the crime scene to test fires from the firearm found on suspect #2. It was determined that the bullet and both cartridge cases were fired from this firearm. The Firearms Unit also identified the shoeprint left at the scene to a shoe found on suspect #2.

The Serology section performed phenolphthalein testing on the bloody items and tested the drinking items for the presence of amylase (a constituent of human saliva.)

The DNA section performed a Polymerase Chain Reaction (PCR) technique to obtain DNA profiles for the above items. The profile from the knife handle matched a known DNA sample from the victim, John Doe, and the blood on the knife matched the known DNA sample from suspect #2. The blood from the shoeprint and the shoe at the scene matched the victim's profile

APPENDIX D. SIMULATION PLOTS





APPENDIX E. EXPERIMENT 2 QUESTION

Standard, state capitals:
Is Philadelphia the capital of Pennsylvania?
a. Yes
b. No
Please rate your level of confidence in your answer on the scale below.

What percentage of people in this survey do you think would answer yes?
_______%

Forecasting:
By April of 2021, will there be more than 300,000 deaths in the United States as a result of COVID-19?
a. Yes
b. No

Please rate your level of confidence in your answer on the scale below.

What percentage of people in this survey do you think would answer yes

Counterfactual:
] "Heavy water" is used to create nuclear isotopes in plutonium which is used in atomic weapons. [2] During World War 2, a Norwegian plant was a main supplier and developer of this heavy water. [3] Norway had been trading this heavy water with Nazis. [4] Allied forces wanted to prevent the Nazis from creating atomic weapons. [5] Resistance groups in Norway were encouraged by allied special forces to destroy these facilities to prevent the German's from acquiring their supply of heavy water. [6] Norwegian resistance tried over many years with no luck. [7] In 1943 – after many failed attempts, Norwegian saboteurs destroyed Germany heavy water supply at a factory Norwegian factory. [8] This greatly reduced the global heavy water supply. [9] In 1944 – the small amount of heavy water supply leftover in Norway was being transported by ship to Germany. [10] A single Norwegian commando snuck on the ship and sunk it [11] with no heavy water supply, German plans to create an atomic weapon were thwarted.

Imagine that the Norwegian resistance was not successful in destroying the factory that produced the heavy water and that you were in a position of making a prediction about various possible outcomes.
Do you believe the Nazis would have been successful at developing a nuclear weapon?

a) Yes
b) No

Please rate your level of confidence in your answer on the scale below.

What percentage of people in this survey do you think would answer yes?
_______%

REFERENCES

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, 295-303.

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-mediated communication and group decision making: A meta-analysis. *Organizational behavior and human decision processes,* 87(1), 156-179.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, 107(3), 797-817.

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22), 10717-10722.

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, 108(36), E625-E625.

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PloS one*, 12(5).

Handgraaf, M. J., Milch, K. F., Appelt, K. C., Schuette, P., Yoskowitz, N. A., & Weber, E. U. (2012). Web-conferencing as a viable method for group decision research. *Judgment and Decision making*, 7(5), 659-668.

Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International journal of man-machine studies*, 34(2), 283-301.

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, 13(4), 322.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22), 9020-9025.

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, 11(3), 279-300.

Nofer, M., & Hinz, O. (2014). Are crowds on the internet wiser than experts? The case of a stock prediction community. *Journal of Business Economics*, 84(3), 303-338.

Olsson, H., de Bruin, W. B., Galesic, M., & Prelec, D. (2019). Harvesting the wisdom of crowds for election predictions using the Bayesian Truth Serum.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306(5695), 462-466.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532-535.

Rutchick, A. M., Ross, B. J., Calvillo, D. P., & Mesick, C. C. (2020). Does the "surprisingly popular" method yield accurate crowdsourced predictions?. *Cognitive research: principles and implications*, 5(1), 1-10.

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: concepts, methods, and recent advancements. *In Disinformation, Misinformation, and Fake News in Social Media* (pp. 1-19). Springer, Cham.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes*, 43(1), 1-28.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in Judge-Advisor decision making. *Organizational behavior and human decision processes*.

Witkowski, J., & Parkes, D. (2012, July). A robust bayesian truth serum for small populations. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 26, No. 1).

Wittenbaum, G. M., Hollingshead, A. B., & Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs*, 71(3), 286-310.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS one*, 10(8).

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, 103(1), 104-120.