2018

# Does the Elicitation Mode Matter? Comparing Different Methods for Eliciting Expert Judgement

Claire Cruickshank

# DOES THE ELICITATION MODE MATTER? COMPARING DIFFERENT METHODS FOR ELICITING EXPERT JUDGEMENT

A Thesis Presented

by

CLAIRE CRUICKSHANK

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

May 2018

INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

**DOES THE ELICITATION MODE MATTER? COMPARING DIFFERENT METHODS FOR ELICITING EXPERT JUDGEMENT**

A Thesis Presented

by

CLAIRE CRUICKSHANK

Approved as to style and content by:

_____
Erin Baker, Chair

_____
Jenna Marquard, Member

_____
Shannon Roberts, Member

_____
Sundar Krishnamurty, Department Head
Mechanical and Industrial Engineering

## ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Erin Baker for the opportunity to undertake this research project and for her support, feedback and guidance. I would also like to thank my committee members Dr. Jenni Marquard and Dr. Shannon Roberts for their feedback and comments on my thesis. I would especially like to thank Dr. Steve Davis for creating the online elicitation and Dr. Karen Jenni for sharing her knowledge and experience conducting face-to-face elicitation interviews. My research was made possible by the Alfred P. Sloan Foundation providing financial support. Finally, I thank the UMass students who volunteered to participate in this research study.

## ABSTRACT

**DOES THE ELICITATION MODE MATTER? COMPARING DIFFERENT METHODS FOR ELICITING EXPERT JUDGEMENT.**

MAY 2018

CLAIRE CRUICKSHANK, B.Sc., UNIVERSITY OF ST ANDREWS

M.S., UNIVERSTITY OF MASSACHUSETTS AMHERST

Directed by: Professor Erin Baker

An expert elicitation is a method of eliciting subjective probability distributions over key parameters from experts. Traditionally an expert elicitation has taken the form of a face-to-face interview; however, interest in using online methods has been growing. This thesis compares two elicitation modes and examines the effectiveness of an interactive online survey compared to a face-to-face interview. Differences in central values, overconfidence, accuracy and satisficing were considered. The results of our analysis indicated that, in instances where the online and face-to-face elicitations were directly comparable, the differences between the modes was not significant. Consequently, a carefully designed online elicitation may be used successfully to obtain accurate forecasts.

## TABLE OF CONTENTS

Page

## LIST OF TABLES

Table Page

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

An expert elicitation is a decision analysis technique used to gather the professional judgements of an individual with expertise in a required field. Decision analysts use this technique when the data required to carry out other statistical approaches are inadequate, unreliable or unavailable.

The subjective probability judgements, gathered from an expert elicitation to characterize the unknown parameter, have been used in a variety of different circumstances to incorporate uncertainty into the decision making process. Expert elicitations have been used in the private sector, for example pharmaceutical companies have used an expert elicitation process to assist executives in deciding how to allocate research and development funds (Sharpe & Keelin, 1998). Also, expert elicitations have been used in the public sector to help guide policy making decisions. One study used structured expert judgements to estimate the global burden of foodborne disease (Aspinall, Cooke, Havelaar, Hoffmann & Hald, 2016). Another study used expert elicitations to characterize the future performance of gas-turbine-based technologies in the electric power sector (Bistline, 2013).

An expert elicitation usually takes the form of a face-to-face interview (F2F) in which an expert is asked to make a series of judgments about the likelihood that an event will occur. In recent years, the traditional face-to-face interview process has been adapted for use as a self-administered online survey (Morgan, 2013, p. 7180). Online elicitation

surveys are less resource intensive, however in this study we are interested in the accuracy of the online elicitation when compared to a face-to-face interview.

The objective of this thesis is to conduct a controlled study and compare these two elicitation modes. Our research examines the interactions between the analyst and expert, and investigates which elicitation mode minimizes the effects of heuristics and biases, while eliciting high quality, accurate probability distributions. As mentioned above, expert elicitations are frequently used to support decision making in the private and public sector and both face-to-face and online modes are used. However there is little research regarding the effect of elicitation mode on the quality of data (Nemet, Anadon & Verdolini, 2017). Insights from our controlled study are intended to inform the design of future expert elicitations which will extend our findings to situations where professional expert judgements of real-world issues are elicited. More specifically, our research will impact the future design of expert elicitations focusing on energy technologies (Baker, 2016).

In order to evaluate the two modes we investigate the level of overconfidence described using the number of surprises and the uncertainty range; the accuracy of the elicited values in estimating the unknown parameter using scoring rules and the detection of the possible use of satisficing by experts during the elicitation.

The rest of the paper is organized as follows. In the remainder of this section, the details of the four research questions are discussed, followed by a literature review. Section 2

describes the elicitation protocol and data analysis methods. Section 3 presents our findings and finally, our conclusion and proposed future work is discussed in section 4.

### 1.1 Research Questions

In this study we investigate four research questions as set out in the original research proposal submitted to the Alfred P. Sloan Foundation. Below we reprint the research questions from the original proposal (Baker, 2016, p. 7). Then we discuss each question in more depth, describe the values and metrics we intend to use to address each research question as well as the expected outcome.

**Table 1.** Research questions adapted from the original proposal (Baker, 2016, p. 7).

| Research Question | Relevant values or metrics | Hypothesis |
|---|---|---|
| Do different modes lead to different central values? | Means of median estimates. | No difference. |
| Which mode results in a larger uncertainty range and less overconfidence? | Uncertainty range and overconfidence. | F2F will have a larger uncertainty range and less overconfidence. |
| Which mode results in more accurate values? | Multiple quantile scoring rule (Jose & Winkler, 2009). | F2F will have more accurate results. |
| Which mode produces satisficing? | Multiple quantile scoring rule (Jose & Winkler, 2009). | F2F will have less satisficing. |

The first research question investigates if different modes lead to different central values. Research has shown that participants' assessment of the median are reasonably accurate

regardless of the shape of the parameter's distribution, symmetrical or highly skewed (Peterson & Miller, 1964). Therefore, we believe the elicitation mode will not impact the accuracy of elicited median values. We hypothesize that we will find no difference in the mean of the elicited median values when we compare the online and face-to-face elicitation modes.

The second research question examines which mode results in less overconfidence. An expert's level of overconfidence is assessed over a series of forecasts. Overconfidence is measured by comparing the proportion of times the observed value falls outside the expert's elicited distributions, referred to as the rate of surprises. In our elicitation, overconfidence is determined by counting the number of times the observed value falls outside the 90% confidence interval. The forecast is perfectly calibrated if the rate of surprises is 10%. If the rate of surprises is above 10%, the judgements have a tendency towards overconfidence (Morgan, 2014). We investigate overconfidence further by using the uncertainty range to indicate the degree of uncertainty. The uncertainty range quantifies the width of the distribution and helps to explain overconfidence. The idea is that a wide uncertainty range is more likely to contain the observed value (Gaba, Tsetlin & Winkler, 2017, p. 4). However, a small uncertainty range, obtained from a narrow distribution, indicates overconfidence as it is more likely that the observed value will fall outside the confidence interval.

The third research question investigates the accuracy of the forecasts. The accuracy of the forecast is quantified using a scoring rule and the choice of scoring rule depends on the

type of assessment used to gather the forecast. In our study, probability assessment and quantile assessment are used. Probability assessment is when a specific value of the parameter is fixed and the cumulative probability associated with each parameter value is assessed (Jose & Winkler, 2009, p. 1287) (Appendix D). However, quantile assessment, is when the specific probability values are fixed and the corresponding parameter values, or the quantiles of the distribution are assessed (Appendix C). A variety of scoring rules have been developed to judge the quality of the forecast. For example, the Brier score, used in weather forecasts, is used for probability assessment (Brier, 1950; cited in Bickel, 2007). Bickel (2007) details three scoring rules used to evaluate probability assessment forecasts: quadratic, spherical and logarithmic scoring rules, and recommends the use of the logarithmic scoring rule for probability assessment.

One important property of a scoring rule is that the expert optimizes their expected score by reporting truthfully their probability assessments (Jose & Winkler, 2009). If this property holds, then the scoring rule is said to be strictly proper. Scoring rules designed for probability assessment are not appropriate for quantile assessment as the scoring rules are no longer strictly proper (Jose & Winkler, 2009). The seven question considered in our analysis use quantile assessment and for that reason we use a linear, strictly proper scoring rule for multiple quantiles detailed by Jose & Winkler (2009, p. 1291).

The fourth research question investigates if satisficing procedures are detected during the elicitation (Simon, 1972). Participants using satisficing procedures do not consider all the possible events, but instead a smaller subset, making the decision when they find the first

solution that meets the criteria. In contrast, an expert elicitation aims to gather carefully considered judgements, where all possible events are taken into consideration before making a decision. In this paper, we investigate if the face-to-face interview produces less satisficing than the online survey. Research suggests that satisficing will be evident in questions appearing near the end of the elicitation, and as a consequence of cognitive fatigue will produce less accurate responses (Krosnick, 1991, p. 214). Therefore, we use the accuracy scores determined using the scoring rule (Jose & Winkler, 2009) to detect satisficing and examine if questions towards the end of the survey are less accurate.

By addressing each research questions we aim to explore whether, and under what circumstances, a self-administered online elicitation offers the same quality of responses compared to the traditional in-person elicitation.

### 1.2 Literature Review

In this section we present a review of literature that examines the effect of survey mode on the quality of participants' responses. First, we look broadly at literature focusing on data gathered from the general public. Then, we briefly review literature regarding some of the cognitive challenges participants encounter during elicitations and the heuristics used by participants to simplify the task. Finally we focus on three articles that combine the results from multiple expert elicitations and use meta-analysis to compare elicitation modes.

Much research focuses on the use of statistical surveys to gather data from the general public including public opinion polls, public health surveys, market research surveys and government surveys. In particular, statistical survey research has investigated the effect of survey mode and suggests that different modes are likely to have an impact on the quality of response data (Bowling, 2005, p. 288). Here, we define the quality of the data in terms of the accuracy of the responses and the absence of response bias. Bowling (2005) describes several advantages in using face-to-face interviews. First of all, participants are required to use less cognitive effort during a face-to-face interview. For example, compared to the self- administered online survey, the face-to-face interview requires as no reading skills. Second, more information may be obtained from a face-to-face interview as the interviewer has the opportunity to encourage longer responses and ask follow-up questions. Also, the presence of the interviewer can enhance the participant's motivation to respond to the survey questions as well as increase the accuracy of the responses. Finally, it is easier for the interviewer to build a rapport with the participant during a face-to-face interview compared to a self-administered online survey because there is visual contact during the interview (Bowling, 2005, p. 288).

However, a disadvantage of the face-to-face interview is social-desirability bias. In other words, the lack of anonymity due to the presence of the interviewer may influence the participant to respond in line with social norms instead of revealing their true beliefs (Bowling, 2005, p. 285). In contrast, the self-administered online survey offers a high level of anonymity. This is one of the main advantages to using self-administered surveys. Research has shown an improvement in the quality of data as participants are

more willing to disclose sensitive information during a self- administered survey compared to a face-to-face interview (Bowling, 2005).

Another way to improve the quality of the response data is to reduce the influence of heuristic procedures. Heuristic procedures, or shortcuts, are used when participants encounter cognitive challenges (Tversky & Kahneman, 1981). However, the use of heuristics produce biased outcomes and errors in judgements (Marquard & Robinson, 2008, p. 7). In this section we briefly summarize: anchoring and adjustment, overconfidence and satisficing.

Anchoring and adjustment is of particular interest during expert elicitations as research has shown this heuristic is present in quantile assessment (Morgan & Henrion, 1990). For example, when a person employs an anchoring and adjustment strategy, they estimate the unknown parameter by starting from some initial value and then adjusting it to obtain the final estimate (Garthwaite, Kadane & O'Hagan, 2005, p. 683). Research has found that often the adjustment is insufficient and the elicited response is biased towards the anchor (Morgan & Henrion, 1990, p. 106). Also, including values in the questions, for example past information will introduce anchoring and influence the forecasts (Marquard & Robinson, 2008, p. 11).

Overconfidence is a common bias seen in expert elicitation. Overconfidence occurs when the participant strongly believes in the accuracy of their predictions (Marquard & Robinson, 2008, p. 13). In the case of expert elicitations, participants' overconfidence in

the accuracy of their predictions results in the observed value falling outside their assessed distribution more often than it should.

Finally, a third concern in expert elicitations, and other surveys, is satisficing. Satisficing is the act of using minimal cognitive effect when responding to survey questions. For example, when a participant gives an initial estimate in response to a question, or in extreme cases responds randomly, as oppose to considering all possible outcomes and finding the optimal response (Krosnick, 1991). Satisficing may happen during a long elicitation and is caused by cognitive fatigue. Research has shown that satisficing is more likely to occur when there is an increase in the difficulty of the task, or a reduction in the participant's ability and motivation to complete the task well (Krosnick, 1991, p. 221). In particular, Krosnick (1991) highlights the difference between weak and strong satisficing. An example of weak satisficing is when a participant gives their initial response as their final answer without carefully considering all the alternatives. Whereas strong satisficing occurs when the expert skims the question and does not fully engage with the material or content but instead gives a superficial response (Krosnick, 1991).

In the remainder of this section we discuss the findings from three articles comparing elicitation modes. The three articles used data from multiple expert elicitations concerning the future cost of energy technology. Some of the expert elicitations were conducted using face-to-face interviews, while others used online elicitation surveys. Researchers were interested to find out if the elicitation mode effected the estimated

future costs of energy technology as well as the degree of uncertainty around the estimates.

Anadon, Nemet & Verdolini (2013) used data from three elicitations. The elicitations gathered data regarding the future cost of nuclear power as a low-carbon power option. Anadon et al. (2013) found no evidence that the elicitation mode had a significant impact on the estimated future costs, however they found on average a lower uncertainty range from the face-to-face elicitations when compared to online. A limitation of their research was the small sample of face-to-face elicitations used in the comparison as well as differences in the background information provided to experts.

The second article by Verdolini, Anadon, Lu & Nemet (2015) used data from five expert elicitations regarding the future cost of photovoltaics, technology that converts light into electricity. Verdolini et al. (2015) found that the elicitation mode did impact expert judgements. Their research found that face-to-face estimates of future costs of photovoltaics were lower and thus more optimistic. Also, their research found that in some cases face-to-face elicitations obtained a larger uncertainty range.

Finally, Nemet, Anadon & Verdolini (2017) used data from 16 elicitations regarding five energy technologies: nuclear, biofuels, bi-electricity, solar and carbon capture. Nemet et al. (2017) found that face-to-face elicitations obtained a larger uncertainty range when compared to online. Nemet et al. (2017) also highlighted that face-to-face elicitations were more costly and time consuming. They concluded that face-to-face elicitations were

more effective at reducing overconfidence and that online elicitations needed further improvements before data gathered using the online mode would be equivalent.

The three articles mentioned have shown that it is likely that the mode effects the elicitation results. Building on the results from the meta-analyses, in this paper we undertake a controlled experiment to evaluate the differences in elicitation modes directly.

# CHAPTER 2

# METHODOLOGY

In this section we give details of our research methods and data analysis.

## 2.1 Elicitation Methodology

In this subsection we describe the design of the online and face-to-face elicitations. First we describe the participants in our study, second the formulation of the elicitation questions, third the face-to-face elicitation protocol and finally we describe how we adapted the face-to-face protocol for use as an online survey.

### 2.1.1 Participants

Our study recruited college students from the University of Massachusetts, Amherst (UMass). Participants were treated as nominal "experts" holding comprehensive and authoritative knowledge of matters of interest to the student population.

Students were invited to respond to recruitment posters placed in various location campus wide including the main university library, campus center and integrated learning building. Also, a notice was placed in the College of Engineering newsletter and emailed to engineering students.

On receipt of expressions of interest, we alternated participants between two groups, placing participants in either the face-to-face interview group or online survey group. Participants' responses were anonymous and pseudonym codes were used to link data.

The individuals assigned to the online survey were sent the web link and instructions on how to access the online elicitation. The online survey was available through the internet and participants chose a convenient time and location to complete the elicitation, for example at home, or at the university library. Individuals assigned to the face-to-face interview group were contacted and a convenient appointment time was agreed. The interviews were conducted in a meeting room in the College of Engineering. During the face-to-face interview, written notes and an audio recording were made. On completion, participants received a thirty dollar gift voucher as payment to compensate for their time.

### 2.1.2 Questions

We prepared twenty questions covering topics of general knowledge and interest to the UMass student population (Appendix A). It was important that our experts, UMass students, would be able to make well-informed judgements. For that reason, our questions were based on the everyday life of students at UMass. For example, we asked questions relating to the UMass library, recreation center and catering services. Also, our questions covered popular culture. For example, we asked participants to predict the opening weekend earnings for an upcoming movie to be shown in the local cinema. We believed college students would have expertise in these topics. However, given that we have a wide variety of questions, not all the students in our study will have particular knowledge about all of the questions. Also, although our experts were college students and not professionals, we believe that the findings from our controlled study will be indicative (Visser, Krosnick, Lavrakas & Kim, 2013, p. 403).

Besides developing questions where UMass students were in a position to make knowledgeable predictions, we formulated questions that met the following requirements. First, our questions were related to unambiguous events or quantities. We took care to construct our questions to avoid ambiguity, confusion and vagueness regarding the unknown parameter. Second, we developed questions that allowed for a valid probability distribution to be elicited (Morgan & Henrion, 1990, p. 50). Finally, we designed questions where the answer was a single observable value that would be measurable in the months after the completion of the elicitations. Our twenty questions are described in Appendix A.

Next, we arranged the questions into the order of appearance in the elicitation. Two different question orders were defined to enable us to investigate the presence of satisficing. Two questions orders allowed for four subgroups: online order 1; online order 2; face-to-face order 1; face-to-face order 2. Each subgroup would contain 20 experts and we believed this would give sufficient statistical power (Appendix B). To determine the question order we first grouped questions into themes, for example questions relating to the UMass library were grouped together. Then questions within the same theme questions were placed in a random order, and the themed groups were randomized to form to different question orders.

### 2.1.3 Face-to-face Elicitation Protocol

The face-to-face elicitations followed a set of procedures, referred to as a protocol. Here we describe the face-to-face protocol used in our study and, in the section that follows we

explain how the face-to-face elicitation protocol was adapted for use as an interactive online survey.

The elicitation protocol provided a systematic approach to elicit subjective probability judgements that was designed to avoid heuristics and biases. Various elicitation protocols have been developed by a number of academic research groups, one of the first was a group of analysts from the Stanford Research Institutes (Morgan & Henrion, 1990, p. 141-145). They developed the Stanford interview process which followed five phases: motivating, structuring, conditioning, encoding and verifying (Morgan & Henrion, 1990, p. 142). Our elicitation protocol followed the five phases of the Stanford interview process.

The first phase of our protocol, referred to as the motivational phase, occurred in the first 5 – 10 minutes of the interview when a rapport with the expert was established. In this opening section of the interview, the analyst presented an overview of the study, then the participant had an opportunity to ask questions and sign the consent form. Also, during this introduction section there was an opportunity to communicate any motivational bias, and in particular the expert had an opportunity to express if their personal situation would influence the elicited judgements.

The second phase of our protocol involved structuring the elicitation questions to avoid ambiguity. We discuss the design of our questions in section 2.1.2. Following on, the third phase, the conditioning phase, focused on avoiding cognitive biases. Several

strategies were used to avoid cognitive biases including the use of follow-up questions. We used follow-up questions to encourage participants to consider the reasons behind their initial judgements as well as to give participants the opportunity to examine all possible outcomes before assessing their judgement. In some instances, on reflection participants altered their subjective probability distributions. Also, we reduced the cognitive challenge of the task, again to avoid cognitive bias, by not pre-determining the units of measurement of the unknown parameter. This strategy reduced the burden of mental calculations and allowed the expert to work in a manner they were comfortable with.

The fourth phase of the protocol involved the encoding of the judgements. This phase occurred during the actual elicitation interview and so in preparation we drafted a script of the conversation between the analyst and the expert. An excerpt from the face-to-face interview is available in Appendix C & D. Every face-to-face interview in our study was unique; the script set out a structure for the interview however the script was adapted during the interview as and when needed.

The interview script was structured to limit the effect of cognitive bias during the encoding phase. To avoid anchoring and adjustment we asked experts to consider the upper and lower limits of the unknown parameter first. We used this strategy to prevent experts anchoring on their best estimate for the median quantile, then adjusting up (or down) to obtain their $95^{th}$ (or $5^{th}$) quantile value. Another strategy we used, this time to reduce overconfidence, was to use interview probes. For example, experts were asked to

explain various scenarios that might cause the observed value to fall below their low estimate (Morgan & Henrion, 1990, p. 144). After encouraging the expert to consider all possible events, some experts decided to alter their judgements. We also prepared pie charts, in place of the standard probability wheel (Morgan & Henrion, 1990, p. 127), as a visual aid to assist with encoding the probability judgements (Figure 1).

The final phase, the verifying phase, asked the expert to reevaluate their judgement. We made some statements based on the elicited distribution to verify the judgements before moving on the next question in the elicitation interview.

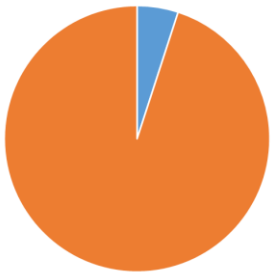
In preparation for the interview we compiled background information relating to each question. We shared a brief summary of the background information and past data with participants at the beginning of each question. Background information was provided to familiarize each expert with the same available knowledge (Morgan, 2013, p. 7179). Also, by carrying out background research, the analyst gained a better understanding of the topic and so was better equipped to challenge and engage the expert during the conversation.

The elicitation interviews were conducted on campus and participants had full access to the internet. We intended for the interview to take around two hours. The actual interviews, not including the introduction section, lasted on average 1 hour 31 minutes, and ranged from one hour four minutes to two hours three minutes (Appendix P). Each participant approached the interview in a different way. Some participants looked up

information using the internet and others used the pen and paper provided to carry out some calculations.

**Figure 1.** Examples of the pie charts used during the face-to-face elicitation.

(a) There is a 1 in 20 chance the spinner will land on blue.



(b) There is a 90% chance the spinner will land on blue.

### 2.1.4 Online Elicitation

The online elicitation survey was administered by Near Zero, a non-profit organization. Near Zero developed software to elicit expert judgements specifically to inform climate and energy policy (Inman & Davis, 2012) and their innovative software was customized for the purposes of our research study.

There are several differences between the face-to-face elicitation protocol and the online elicitation. First, the question wording and approach was adapted slightly to take advantage of the software's interactive graphical features. However, it was important that both survey modes contained the same background information and definitions to allow for a fair comparison. Second, the presentation of the background information and definitions differed. In the online elicitation information was provided in rollovers. In other words, when the participant rolled the mouse curser over the highlighted text,

additional information was displayed (see Figure 2). Rollovers were used to avoid overwhelming participants with large sections of written instructions.

Similar to the face-to-face elicitation, the online software also gathered qualitative information. Open questions were included in the online elicitation, giving participants an opportunity to type a written response. Participants' written comments provided valuable insights into the participants thinking and allowed for transparency. In our elicitation, participants responded to 94% of our open questions; only 4% of the responses were either "don't know", "unsure" or "NA" and 2% of the open questions were left blank.

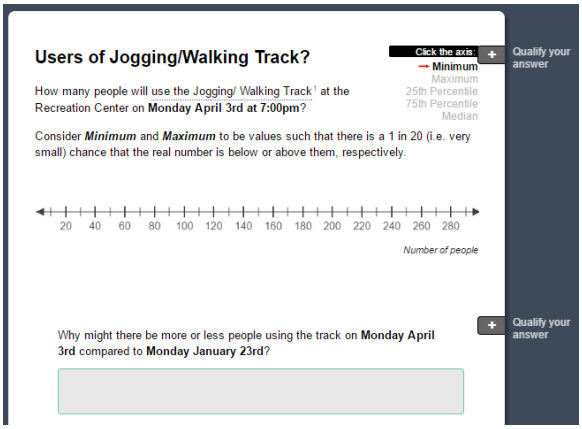
Questions were presented in sequential order, one question per webpage. Three different widgets were used: box-and-whisker, time trend graph and direct entry (Figure 3). The box-and whisker widget and time trend graph were used for quantile assessment; whereas the direct entry widget was used for probability assessment.

The box-and-whisker widget was used most often: in eighteen out of the twenty questions. One feature of the box-and-whisker widget was that the instructions were shown on the top right hand corner of the webpage (Figure 2). Participants were asked for a minimum value ($5^{th}$ percentile) and maximum value ($95^{th}$ percentile), $25^{th}$, $75^{th}$ and 50th percentile. The red arrow appeared in the instruction box to indicate the requested percentile. Instructions were concise and the widget design intuitive. After the participant selected their percentile judgements, a box-and-whisker plot was displayed. If the participant needed to make changes to their values they were able to click on the box-

and-whisker plot and make the required adjustments. However, a weakness of this method was that participants were presented with an initial high and low value as a number line was displayed on screen. Although the number line changed as participants selected their values, the inclusion of the initial high and low values in the question almost certainly anchored answers (Marquard & Robinson, 2008, p. 11). Nevertheless, we used this method as this was the best practice in online elicitations.

**Figure 2.** Illustration of box-and-whisker widget.

(a) Participants clicked on the number line and wrote comments in the text box.




(b) Background information and definitions were presented as rollovers.

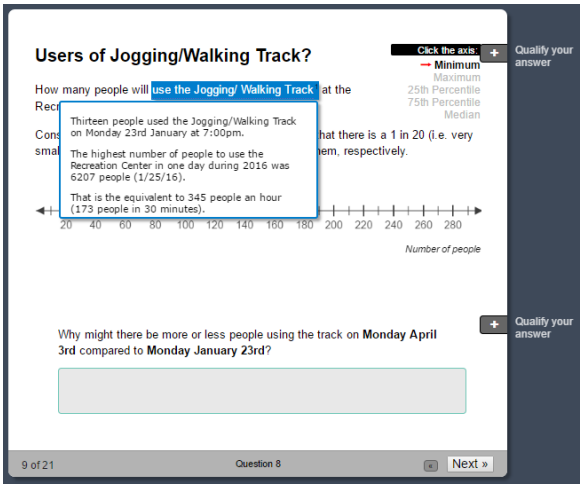
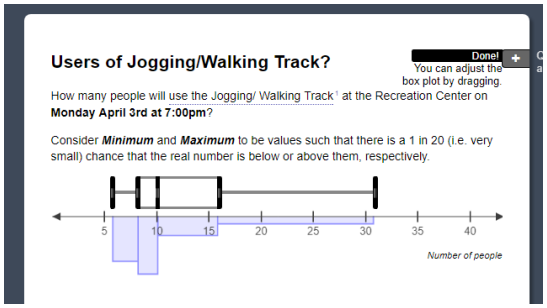
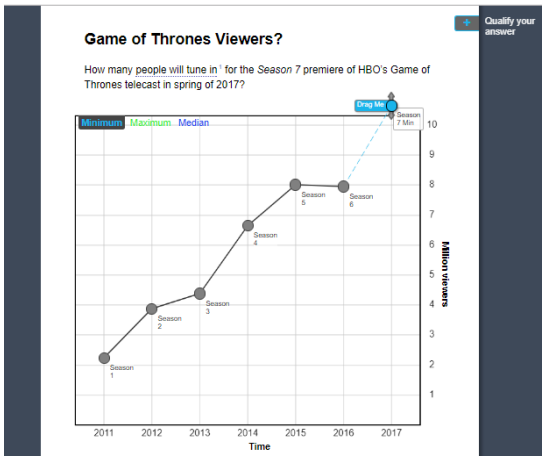


**Figure 3.** Examples of widgets

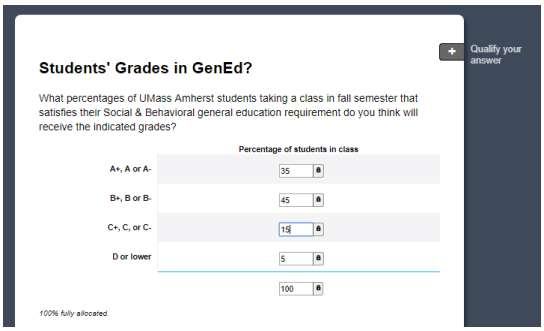
(a) Box-and-whisker widget




(b) Time trend widget




(c) Direct entry widget where participants type percentages directly.




## 2.2 Data Analysis Methodology

In this section we discuss the methods of data analysis. In section 2.2.1, we describe how data was prepared for analysis. Next, we detail the mathematical notation used in the paper. In section 2.2.3, we describe how we constructed aggregated distributions and combined the judgements of multiple experts. In section 2.2.4, we describe how we compared the central values. In section 2.2.5 and 2.2.6 we examine the level of overconfidence by considering the uncertainty range and rate of surprises. In section 2.2.7 we define the scoring rule, then following on, in section 2.2.8 we describe the

question-by-question approach used to examine the accuracy of the judgements. In section 2.2.9 we describe how scores were normalized and then in section 2.2.10 we detail how we combined normalized scores across questions. Finally, in section 2.2.11 we describe our methods used to investigate the presence of satisficing.

### 2.2.1 Preparing Data for Analysis

On completion of the elicitation interviews and online surveys we had a data set consisting of subjective probability distributions. Specifically the $5^{th}$, $50^{th}$ and $95^{th}$ percentile values from the cumulative distribution corresponding to each question in our elicitation, with every expert providing one distribution per question. We took steps to improve the quality of the data in order to be sure that our conclusions were based on valid reasons, as opposed to a result of a mistake with data entry or missing data values (Osborne, 2008, p. 198). We identified errors in our survey data and if inconsistencies appeared, we removed the data.

We identified the following problems with some of the online questions: the observed value occurred outside the pre-determined range shown on the online survey; the pre-determined units of measurements on the online survey were not the most appropriate choice; the online software did not function correctly. Based on these inconsistencies, we identified seven questions from our elicitation consistent between face-to-face and online. We grouped these questions together and throughout this paper we refer to this group as group 1 (Appendix E). Our results are reported for group 1.

The remaining questions, not included in our analysis, were categorized into three groups. Group 2 included four questions in which the observed value occurred outside the pre-determined range shown on the online survey. Questions in group 2 were not comparable between face-to-face and online because the participants completing the face-to-face elicitation were not shown a pre-determined range. Instead, during the face-to-face elicitation participants were asked to set the high and low estimate. Therefore, we believe the heuristic of anchoring and adjustment influenced the online elicited distributions.

Group 3 included four questions in which the pre-determined unit of measurements on the online survey were not the most appropriate choice. Specifically, the online survey measured in minutes however seconds would have been a more appropriate choice. The final group, group 4, included questions where the elicitation data was not complete for a variety of different reasons including the online software did not function correctly.

We note here that in all of the questions excluded from our analysis (group 2 and 3), the face-to-face was more accurate than the online at the 6% level (Appendix Q and R).

### 2.2.2 Notation

We use the following notation throughout the paper.

$i$: Expert, $i \in \{1, 2, \dots, 73\}$.

$j$: Survey mode, $j \in \{1, 2\}$ where mode 1 ($j = 1$) is the online elicitation and mode 2 ($j = 2$) is the face-to-face elicitation.

$k$: Question, $k \in \{1, 2, \dots, 7\}$. We use seven questions to compare modes.

$N_k$: Number of forecasts elicited for question $k$.

$n_{jk}$: Number of forecasts elicited from survey mode $j$ for question $k$.

$F_j$: Number of forecasts elicited from mode $j$ across questions: $F_j = \sum_{k=1}^{7} n_{jk}$.

$a_y$: $a^{\text{th}}$ percentile. $a_1$ is the 5$^{\text{th}}$ percentile ($a_1 = 5$), $a_2$ is the 50$^{\text{th}}$ percentile ($a_2 = 50$) and $a_3$ is the 95$^{\text{th}}$ percentile ($a_3 = 95$).

$q_{a_y\, ijk}$: The value of the elicited $a_y{}^{\text{th}}$ percentile for expert $i$, survey mode $j$, question $k$.

$T_k$: Observed value (true value) for question $k$.

$\bar{x}_{jk}$: The average median estimate for survey mode $j$, question $k$.

$M_{jk}$: The average score for survey mode $j$, question $k$.

### 2.2.3 Aggregate Individual Expert Distributions

For each question, we aggregated individual experts' distributions into a cumulative distribution for the online experts, and a cumulative distribution for the face-face experts. We assumed experts' beliefs were independent, although Usher and Strachan (2013) highlighted that the assumption of independence was strong and unlikely to exist in real life. In other words, our experts were likely to base their judgements on similar experiences and background knowledge and hence it was likely our experts were correlated. Nonetheless, we assumed independence and that each expert was equally credible (Clemen & Winkler, 1999). To combine the experts' judgements we used equal weight aggregation (Lichtendahl, Grushka-Cockayne & Winkler, 2013) and we computed the arithmetic mean of the elicited percentile values as follows:

$$aggregate\ percentile_{a_y jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} q_{a_y\, ijk}$$

The aggregated individual expert distributions are displayed in Appendix G (also see, Appendix Q for aggregated distributions of questions in group 2 and 3).

### 2.2.4 Comparison of the Central Values

In this section we investigate if different modes led to different central values. We considered each question separately and used a question-by-question approach to evaluate the effect of elicitation modes on elicited median values.

We completed our analysis in three stages. The first stage involved computing descriptive statistics to summarize the elicited median values. Also, we used a variety of graphs: histograms, boxplots and Q-Q plots (quantile-quantile plot) (Wilk & Gnanadesikan, 1968) to identify patterns in our data and inspect the underlying distribution. In particular, we used the Q-Q plots to check for normality (Ghasemi & Zahediasl, 2012) and then the Shapiro-Wilk test to verify the normality assumption (Shapiro & Wilk, 1965, p. 593).

In the second stage of our analysis, we further examined the shape of the underlying distribution by calculating skewness and kurtosis then identifying the outliers of the distribution (Barton & Peat, 2014, p24). The skewness of a data population was defined as follows:

$$\gamma_1 = \frac{\mu'''}{\mu''^{\frac{3}{2}}}$$

where $\mu''$ represents the second central moment and $\mu'''$ represents the third central moment (Cramér, 1946; see also Joanes & Gill, 1998; Revelle, 2017, p. 228).

The Pearson's measure of kurtosis was defined as follows:

$$\gamma_2 = \frac{\mu''''}{\mu''^{\,2} - 3}$$

where $\mu''$ represents the second central moment and $\mu''''$ represents the fourth central moment (Cramér, 1946; see also Joanes & Gill, 1998; Revelle, 2017, p. 228).

In the third stage of our analysis we used inferential statistics to investigate if the elicited median responses from the face-to-face mode were significantly different from the elicited median values from the online mode. For a given question ($k$), we tested the null hypothesis: the mean of the online elicited median values was equal to the mean of the face-to-face elicited median values. We tested against the alternative hypothesis: the mean of the online elicited median values was not equal to the mean of the mean of the face-to-face elicited median values.

For each question, we used an independent two sample $t$ -tests to compare the means of the elicited median values. We assumed the two samples were independent; the variances were unknown and unequal; the sample sizes were unequal and large. We defined a large sample size to be greater than 30 samples (Mann, 2007, p. 458; see also Ghasemi & Zahediasl, 2012).

We estimated the population mean using the sample mean ( $\bar{x}_{jk}$) and we defined the sample mean of the elicited median values as follows:

$$\bar{x}_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} q_{a_2\, ijk}$$

Next, we computed the standard deviation ($s_{jk}$) of the sample, defined as follows:

$$s_{jk} = \sqrt{\frac{\sum_1^{n_{jk}} (q_{a_2\, ijk} - \bar{x}_{jk})^{\,2}}{n_{jk} - 1}}$$

We used the Satterthwaite Approximation to measure the standard error. The Satterthwaite Approximation takes into account of the unequal variances and unequal sample sizes by computing a weighted average of the standard errors ($SE_k$) We defined the standard error using the following equation (Moser, Stevens & Watts, 1989, p. 3964).

$$SE_k = \sqrt{\frac{{s_{1k}}^2}{n_{1k}} + \frac{{s_{2k}}^2}{n_{2k}}}$$

Then, we computed the test statistic, the Welch's $t$-test, defined by the following equation (Mann, 2007, p. 460; also see Moser, Stevens & Watts, 1989):

$$t = \frac{(\bar{x}_{1k} - \bar{x}_{2k})}{SE_k}$$

Finally, the degrees of freedom ($df$) was approximated using the Welch-Satterthwaite equation defined as follows (Mann, 2007, p. 458):

$$df = \frac{\left(\frac{{s_{1k}}^2}{n_{1k}} + \frac{{s_{2k}}^2}{n_{2k}}\right)^2}{\frac{\left(\frac{{s_{1k}}^2}{n_{1k}}\right)^2}{(n_{1k}-1)} + \frac{\left(\frac{{s_{2k}}^2}{n_{2k}}\right)^2}{(n_{2k}-1)}}$$

In cases where that data are normally distributed, we calculate the effect size to determine the magnitude of the difference between the online and face-to-face elicitation. We used Hedge's $g$ statistics to estimate the population effect size, defined as follows (Barton & Peat, 2014, p. 57):

$$g = \frac{\bar{x}_{1k} - \bar{x}_{2k}}{\sqrt{\frac{\left((n_{1k}-1){s_{1k}}^2 + (n_{2k}-1){s_{2k}}^2\right)}{n_{1k} + n_{2k} - 2}}}$$

### 2.2.5 Comparison of the Uncertainty Range

In this section we investigate which mode resulted in a larger uncertainty range. The uncertainty range is a measure of the percentage variation from each expert's median estimate. We defined the uncertainty range as the difference between the 95$^{th}$ and the 5$^{th}$ percentile of the unknown parameter, normalized by the median (Anadon, Nemet & Verdolini, 2013, p. 3; also see Verdolini, Anadon, Lu & Nemet, 2015). The following formula was used to calculate the normalized uncertainty range ($NUR_{ijk}$):

$$NUR_{ijk} = \frac{q_{a_3\,ijk} - q_{a_1\,ijk}}{q_{a_2\,ijk}}$$

To assess the effect of elicitation mode on the width of the distribution, we computed the normalized uncertainty range for every expert ($i$), across each question ($k$). Next, we

used inferential statistics and tested the null hypothesis that the mean normalized uncertainty range from the face-to-face elicitation was less than the mean normalized uncertainty range from the online. We estimated the population mean using the average normalized uncertainty range ($ANUR_j$) defined as follows:

$$ANUR_j = \frac{1}{F_j} \sum_{k=1}^{7} \sum_{i}^{n_{jk}} NUR_{ijk}$$

### 2.2.6 Comparison of the Rate of Surprises

In this section we investigate which mode led to less overconfidence by examining the numbers of surprises (Budescu and Du, 2007, p. 1732). We defined a surprise ($c_{ijk}$) as the event that the observed value that lies outside the 5-95 range:

$$c_{ijk} = \begin{cases} 0 & if \ \ q_{a_1\, ijk} < T_k < q_{a_3\, ijk} \\ 1 & otherwise \end{cases}$$

To evaluate the effect of elicitation mode on the level of overconfidence, we totaled the number of surprises across questions. Then, we used inferential statistics to test if the face-to-face elicitations resulted in a lower proportion of surprises than the online.

The null hypothesis was as follows: the proportion of surprises for online participants was not greater than the proportion of surprises for face-to-face participants. We tested against the alternative hypothesis: the proportion of surprises for online participants was greater than that for face-to-face.

We defined $\hat{p}_j$ as the proportion of surprises in our sample for a given mode ($j$) as follows:

$$\widehat{p_j} = \frac{1}{F_j} \sum_{k=1}^{7} \left( \sum_{i=1}^{n_{jk}} c_{ijk} \right)$$

We defined the pooled estimate ($\bar{p}$) as follows: (Mann, 2007, p. 476):

$$\bar{p} = \frac{\widehat{p_1} F_1 + \widehat{p_2} F_2}{F_1 + F_2}$$

Standard Error ($SE$)was defined by the following equation (Mann, 2007, p. 476):

$$SE = \sqrt{\frac{\bar{p}(1 - \bar{p})}{F_1} + \frac{\bar{p}(1 - \bar{p})}{F_2}}$$

The test statistics for two independent proportions was defined by the following equation (Mann, 2007, p. 476):

$$z = \frac{(\widehat{p_1} - \widehat{p_2})}{SE}$$

### 2.2.7 Accuracy of Forecast

We used the multiple quantile scoring rule to assess the accuracy of the forecasts. The multiple quantile scoring rule combines the three assessed quantiles and the observed value to give a core accuracy score. We defined the scoring rule for multiple quantile assessment as follows (Jose & Winkler, 2009, p. 1290; also see Grushka-Cockayne, Jose & Lichtendahl, 2017, p. 1122):

$$S\left(q_{a_y\,ijk}, T_k\right) = \begin{cases} \sum_{y=1}^{3} \left| T_k - q_{a_y\,ijk} \right| (a_y) & for\ T_k \geq q_{a_y ijk} \\ \sum_{y=1}^{3} \left| T_k - q_{a_y\,ijk} \right| (100 - a_y) & for\ T_k < q_{a_y ijk} \end{cases}$$

$$S\left(q_{a_y\,ijk}, T_k\right) \in [0, \infty)$$

The multiple quantile scoring rule has the following properties. First, it only takes positive values and has no upper bound. Second, the multiple quantile scoring rule is strictly proper (Jose & Winkler, 2009, p. 1291, proposition 4.1) and a low core score indicates a more accurate forecast. Finally, the multiple quantile scoring rule is scale-dependent, and therefore the core scores are expressed in the units of the assessed quantiles.

### 2.2.8 Comparison of the Accuracy of the Forecasts

In this section we investigate if the face-to-face elicitation produced more accurate forecasts. We considered each question separately and used a question-by-question approach to evaluate the effect of elicitation mode on the accuracy of the judgements. We completed our analysis in three stages.

In the first stage of our analysis, we assigned a core score to each forecast. Then we summarized the core scores data using descriptive statistics. Also, we used a variety of graphs: histograms, boxplots and Q-Q plots (quantile-quantile plot) (Wilk & Gnanadesikan, 1968) to identify patterns in our data and inspect the underlying distribution. As before, we used Q-Q plots to check for normality (Ghasemi & Zahediasl,

2012) and then the Shapiro-Wilk test to verify the normality assumption (Shapiro & Wilk, 1965, p. 593).

In the second stage of our analysis, we further examined the shape of the underlying distribution by calculating skewness and kurtosis, defined in section 2.2.4. Then we identified any outliers (Barton & Peat, 2014, p. 24).

In the third stage of our analysis we used inferential statistics to investigate if the face-to-face mode produced more accurate forecasts. For a given question ($k$), we tested the null hypothesis: the mean core score from the online elicitation was not greater than that from the face-to-face. We tested against the alternative hypothesis: the mean core score from the online elicitation was greater than that from the face-to-face. In cases where the online mode had better (lower) core scores on average, we tested the opposite hypothesis: that the online mode is more accurate than the face-to-face.

We used independent two sample $t$-tests to compare average core scores. We assumed the two samples were independent; the variances were unknown and unequal; the sample sizes were unequal and large. Again, we defined a large sample size to be greater than 30 samples (Mann, 2007, p. 458; see also Ghasemi & Zahediasl, 2012).

We estimated the population mean using the sample mean $M_{jk}$ defined as follows:

$$M_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} S\left(q_{a_y\, ijk}, T_k\right)$$

Next, we computed the standard deviation ($s_{jk}$) of the sample, defined as follows:

$$s_{jk} = \sqrt{\frac{\sum_1^{n_{jk}} \left( S\left(q_{a_y\, ijk}, T_k\right) - M_{jk} \right)^2}{n_{jk} - 1}}$$

We used Satterthwaite Approximation to represent the standard error, defined as:

$$SE_k = \sqrt{\frac{{s_{1k}}^2}{n_{1k}} + \frac{{s_{2k}}^2}{n_{2k}}}$$

Then, we computed the test statistics. We used the Welch's $t$ statistics for unequal variances, defined as follows (Mann, 2007, p. 460):

$$t = \frac{(M_{1k} - M_{2k})}{SE_k}$$

Finally, we computed the degrees of freedom defined as: (Mann, 2007, p. 458):

$$df = \frac{\left(\frac{{s_{1k}}^2}{n_{1k}} + \frac{{s_{2k}}^2}{n_{2k}}\right)^2}{\frac{\left(\frac{{s_{1k}}^2}{n_{1k}}\right)^2}{(n_{1k} - 1)} + \frac{\left(\frac{{s_{2k}}^2}{n_{2k}}\right)^2}{(n_{2k} - 1)}}$$

### 2.2.9 Normalizing Scores

In this section, we discuss how we normalized the accuracy score to allow us to make a fair comparison across questions. As mentioned before in section 2.2.7, the multiple

quantile scoring rule is based directly on the scale of the unknown parameter; it is scale-dependent (Jose, 2017; also Hyndman and Koehler, 2006; Patton 2011). In other words, if the elicitation questions are expressed in the same units, a valid comparison across question can be made. However if, like in our elicitation, questions are expressed in different units, then this presents a challenge. In our case, we were unable to use the cores scores in their current form to make a valid comparison because the questions were expressed in variety of units including: seconds, millions of dollars, and degrees Fahrenheit (Hyndman and Koehler, 2006, p. 682; also see Jose, 2017, p. 200).

To overcome this challenge, we adjusted the core scores to a notionally common scale and we used a scale-independent, linear transformation to normalize the core scores. In particular, we used linear transformations so to preserve the strictly proper scoring property (Toda, 1963; also see Bickel, 2007; Jose, 2009, p.1295). We defined the linear transformation function as follows:

$$B_{bijk} = \ e \ + \ f \ A_{ijk}$$

The notation used:

$A_{ijk}$: Core score for expert $(i)$, survey mode $(j)$ and question $(k)$

$B_{bijk}$: Normalized score using linear transformation $(b)$, for expert $(i)$, survey mode $(j)$ and question $(k)$

$e$ : Normalization constant and $y$-intercept of the linear transformation function.

$f$: Normalization constant and gradient of the linear transformation function.

Next, we examined the effect of two linear transformations on the core scores to establish the most appropriate transformation. First, we considered normalizing core scores by dividing by the mean. Then second, we considered normalizing using a method referred to as min-max normalization. Both transformations are defined below.

The first linear transformation ($b = 1$) was computed by dividing by the arithmetic mean of the elicited values ($M_k$):

$$M_k = \frac{1}{N_k} \sum_{j=1}^{2} \sum_{i} A_{ijk}$$

We define the linear transformation as follows:

$$B_{1\,ijk} = \frac{A_{ijk}}{M_k}, \qquad B_{1\,ijk} \in [0, \infty)$$

The normalization constants for this linear transformation are: $e = 0$ and $f = \frac{1}{M_k}$.

The second linear transformation ($b = 2$), min-max normalization was defined as follows:

$$B_{2\,ijk} = \frac{A_{ijk} - \min_{ij} A_{ijk}}{\max_{ij} A_{ijk} - \min_{ij} A_{ijk}}, \qquad B_{2\,ijk} \in [0, 1]$$

In this linear transformation the normalization constants are:

$$e = \frac{-\min_{ij} A_{ijk}}{\max_{ij} A_{ijk} - \min_{ij} A_{ijk}}$$

$$f = \frac{1}{\max_{ij} A_{ijk} - \min_{ij} A_{ijk}}$$

Bickel (2007) recognized that normalizing scores was challenging because the core scores have no upper bound and although, in his paper Bickel was working with probability assessments, he used an approach equivalent to the min-max transformation to normalize the accuracy scores.

Following on, we evaluated the performance of the two both transformations. First we used scatter plots to examine the effect of the transformation. Second, we examined the shape of the underlying distribution and examined the normalized mean, standard deviations, skew and kurtosis.

### 2.2.10 Comparison of the Accuracy of the Forecasts across Questions

In this section, we compare the accuracy scores across questions. Our analysis was carried out in three stages. First, we aggregated the normalized scores and assigned a single accuracy score to each mode. To do so, we used equal weight averaging defined as:

$$M_j = \frac{1}{F_j} \sum_{k=1}^{7} \sum_{i}^{n_{jk}} B_{bijk}$$

where $M_j$ represents the average normalized score for survey mode $j$.

After investigating the normality assumption, we carried out an independent two sample $t$–test. We estimated the population mean using the sample mean $M_j$ (defined above) and we computed the standard deviation ($s_j$) to be the standard deviation of the sample defined as follows:

$$s_j = \sqrt{\frac{\sum_{k=1}^{7}\sum_{1}^{n_{jk}}(B_{b\,ijk} - M_j)^2}{F_j - 1}}$$

We used Satterthwaite Approximation to represent the standard error, defined as:

$$SE = \sqrt{\frac{{s_1}^2}{F_1} + \frac{{s_2}^2}{F_2}}$$

Then we computed the test statistics, Welch's $t$ statistics, defined by the following equation (Mann, 2007, p. 460):

$$t = \frac{(M_1 - M_2)}{SE}$$

Finally we computed the degrees of freedom defined by the following equation (Mann, 2007, p. 458):

$$df = \frac{\left(\frac{{s_1}^2}{F_1} + \frac{{s_2}^2}{F_2}\right)^2}{\frac{\left(\frac{{s_1}^2}{F_1}\right)^2}{(F_1 - 1)} + \frac{\left(\frac{{s_2}^2}{F_2}\right)^2}{(F_2 - 1)}}$$

### 2.2.11 Comparison of the Presence of Satisficing

In this section we investigate which mode produced satisficing. Satisficing is where the participant completes the survey quickly by doing enough to "satisfy" the survey. We hypothesized that satisficing would be evident towards the end of the elicitation when the experts would be fatigued and more inclined to use cognitive shortcuts. We explored if the experts interactions with the analyst during the face-to-face interview resulted in less satisficing. To do this, we considered if a question appearing early in the elicitation produced a more accurate forecast than when the same question appeared near the end of the elicitation.

To carry out the analysis, we used the library elevator question ($k = 1$) to make the comparison as the question appeared near the beginning of the elicitation (question 3) and near the end (question 18) in the two orders. Participants from both survey modes were asked to make a judgement on the time taken to travel in the elevator to the 23$^{rd}$ floor of the library. We carried out an independent two sample $t$-test to find out if the question appearing early in the elicitation obtained a lower core score. If this is true, then this would indicate the use of satisficing procedures towards the end of the elicitation.

# CHAPTER 3

# RESULTS

The subjects in this experiment were 73 undergraduate and graduate students; 39 individuals completed the face-to-face interview and 34 individuals completed the online survey.

In this section we detail our results. In section 3.1 we address our first research question and compare the central values. In section 3.2 and 3.3 we address our second research question to find out which mode results in less overconfidence. We investigate the level of overconfidence by comparing the uncertainty range and the rate of surprises. Following on, we examine which mode results in more accurate values. We use a scoring rule to assess the accuracy of the modes. Details of the scoring rule are described in section 3.4. In section 3.5 we use a question-by-question approach to compare the accuracy of the forecasts. Next, in section 3.6 we examine how to normalize the core score before, in section 3.8 we compare the accuracy of the forecasts across questions. Finally in section 3.7, we consider our fourth research question and explore which mode produces satisficing.

## 3.1 Comparison of the Central Values Results

Here we investigate the first research question: do different modes lead to different central values? We hypothesized that there would be no difference in the central values. This is largely confirmed by our results as we found a significant difference in only two

out of the seven questions. Table 2 below summarizes the results from independent sample $t$ -test.

We found a significant difference in the elicited median values in two questions: the Game of Thrones question and YouTube question. In the Game of Thrones question, the more accurate forecast were obtained from the online elicitation. This result was expected because the online elicitation used a time trend widget (Figure 3) and we speculated that the use of interactive software would obtain more accurate judgements.

Regarding the second question, the YouTube question, we found that the face-to-face elicitation obtained a more accurate forecast. Again, this result was expected because the YouTube question involved calculations. In this situation, the interaction with the analyst was an advantage of the face-to-face mode as part of the analyst's role was to ask further questions and to encourage the expert to think beyond their initial best guess.

After completing the three stages of our analysis, we found that the underlying distributions of both questions, were non-normal. However, we argued that the $t$-test was valid because the sample sizes are large ($n > 30$). In summary, we found that in five out of seven questions no statistically significant difference between the mean elicited median value. Therefore, the results support the idea that there was no difference in the central values (Appendix I, G and H).

**Table 2.** Did different modes led to different central values? Summary of results from independent samples tests comparing the online mean median estimate with the face-to-face.

| $k$ | Question | Significant difference? | Online $M$ ($SD$) | F2F $M$ ($SD$) | $t$ | $df$ | $p$ (two-sided) | Normality assumption holds? (g***) |
|---|---|---|---|---|---|---|---|---|
| 1 | Library elevator | No | 78.80 (65.51) | 99.92 (32.82) | $-1.67$ | 45 | .101 | No |
| 2 | Hip hop class | No | 33.85 (14.22) | 38.56 (10.41) | $-1.60$ | 59 | .1157 | Yes ($-0.38$) |
| 3 | Basketball attendance | No | 3048 (1493) | 3173 (739) | $-0.44$ | 45 | .6646 | Yes ($-0.11$) |
| 4 | Game of Thrones | Yes* | 9.22 (1.17) | 8.07 (0.70) | 4.98 | 50 | $<$ .001 | Online yes |
| 5 | YouTube | Yes** | 42.77 (37.47) | 24.18 (7.04) | 2.85 | 35 | .007301 | No |
| 6 | Opening weekend | No | 113.28 (46.29) | 105.46 (19.18) | 0.92 | 42 | .3637 | Yes (0.22) |
| 7 | High temperature | No | 60.15 (10.11) | 56.18 (7.62) | 1.81 | 56 | .07606 | F2F yes |

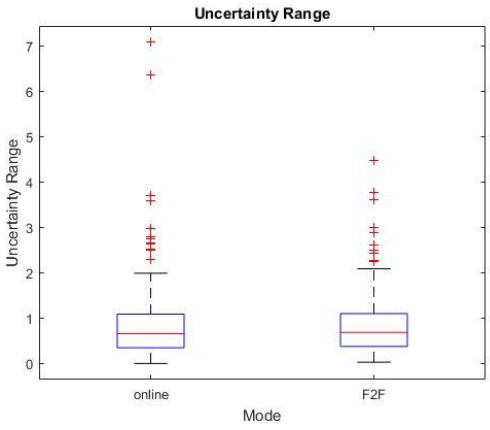
*Online elicitation obtains more accurate forecasts (based on scoring rule).
** F2F elicitation obtains more accurate forecasts (based on scoring rule).
*** g denotes Hedges measure of effect size with $|g| < 0.2$ "negligible", $|g| < 0.5$ "small"

### 3.2 Comparison of the Uncertainty Range Results

We investigate which mode led to a higher uncertainty range. The uncertainty range describes the interval width and indicates the degree of uncertainty. A wide interval will give a higher uncertainty range and indicate less overconfidence. The normalized uncertainty range from the online elicitation was on average 86% of the experts' median estimate, and 85% in the case of the face-to-face.

We tested if the face-to-face uncertainty range was significantly higher compared to online. An independent two sample $t$-test was conducted to compare the mean uncertainty range. The boxplots (Figure 4) indicated that the underlying distributions were positively skewed and the normality assumption does not hold (Appendix J). Contrary to our hypothesis, and while not statistically significant, the online elicitation showed a slightly higher mean uncertainty range. Results indicated however that the mean uncertainty range for the online elicitation ($M = 0.86, SD = 0.82, n = 267$), was not significantly higher at $\alpha = .05$ than the face-to-face elicitation ($M = 0.85, SD = 0.7, n = 309$), $t(524) = 0.17, p = .5674$, one-tailed.

**Figure 4.** Boxplot comparing the effect of elicitation mode on the uncertainty range.

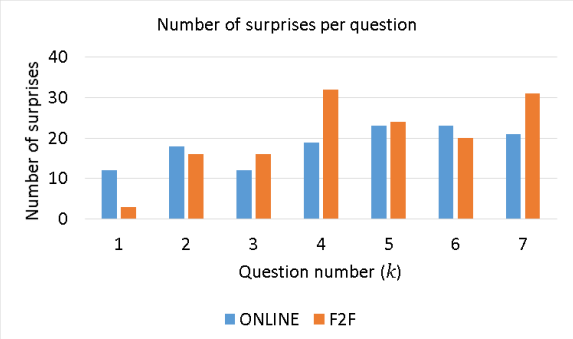


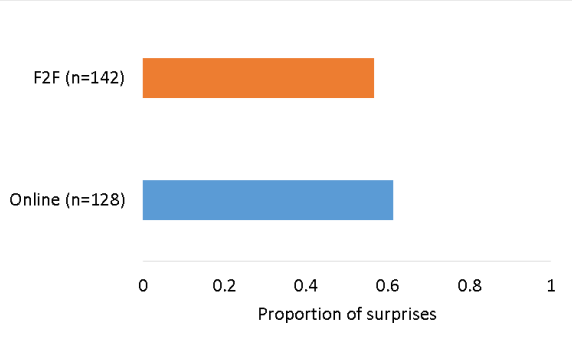
### 3.3 Comparison of the Rate of Surprises Results

We investigate which survey mode resulted in less overconfidence in terms of the rate of surprises. We computed the frequency of surprises by counting the number of times the observed value fell outside the 90% confidence interval. The forecast was perfectly calibrated if the rate of surprises was 10% (p = 0.10). However, the online and face-to-face elicitations produced considerably higher rate of surprises: 61% and 56% respectively (Figure 5). Research has shown that quantile assessments produced high levels of overconfidence where experts' probability distributions were too narrow and rate of surprises greater than 10% (Garthwaite et al., 2005, p. 685).

After completing our analysis, we found that the proportion of surprises from the online elicitation (p = 0.61) was not significantly higher than the face-to-face elicitation (p = 0.56), $z = 0.90, p = .4086$, one-tailed. Therefore, the evidence was not sufficient to state that the face-to-face responses were less overconfident (Appendix K).

**Figure 5.** Comparing the effect of elicitation mode on the level of overconfidence.
(a) Frequency surprises. (b) Proportion of surprise



### 3.4 Accuracy of the Forecasts Results

We investigate which mode results in more accurate forecasts using the multiple quantile scoring rule detailed in section 2.2.7. In this section we illustrate an application of the multiple quantile scoring rule.

In this illustration, participants were asked to forecast how many people would participate in the cardio hip hop class at 7:30pm on Monday 3rd April. On April 3rd, we observed 11 people participating in the class. Then, if an expert expressed their forecast of the ($5^{th}$, $50^{th}$, $95^{th}$) percentile as: (8 people, 30 people, 40 people), we compute their core score as follows:

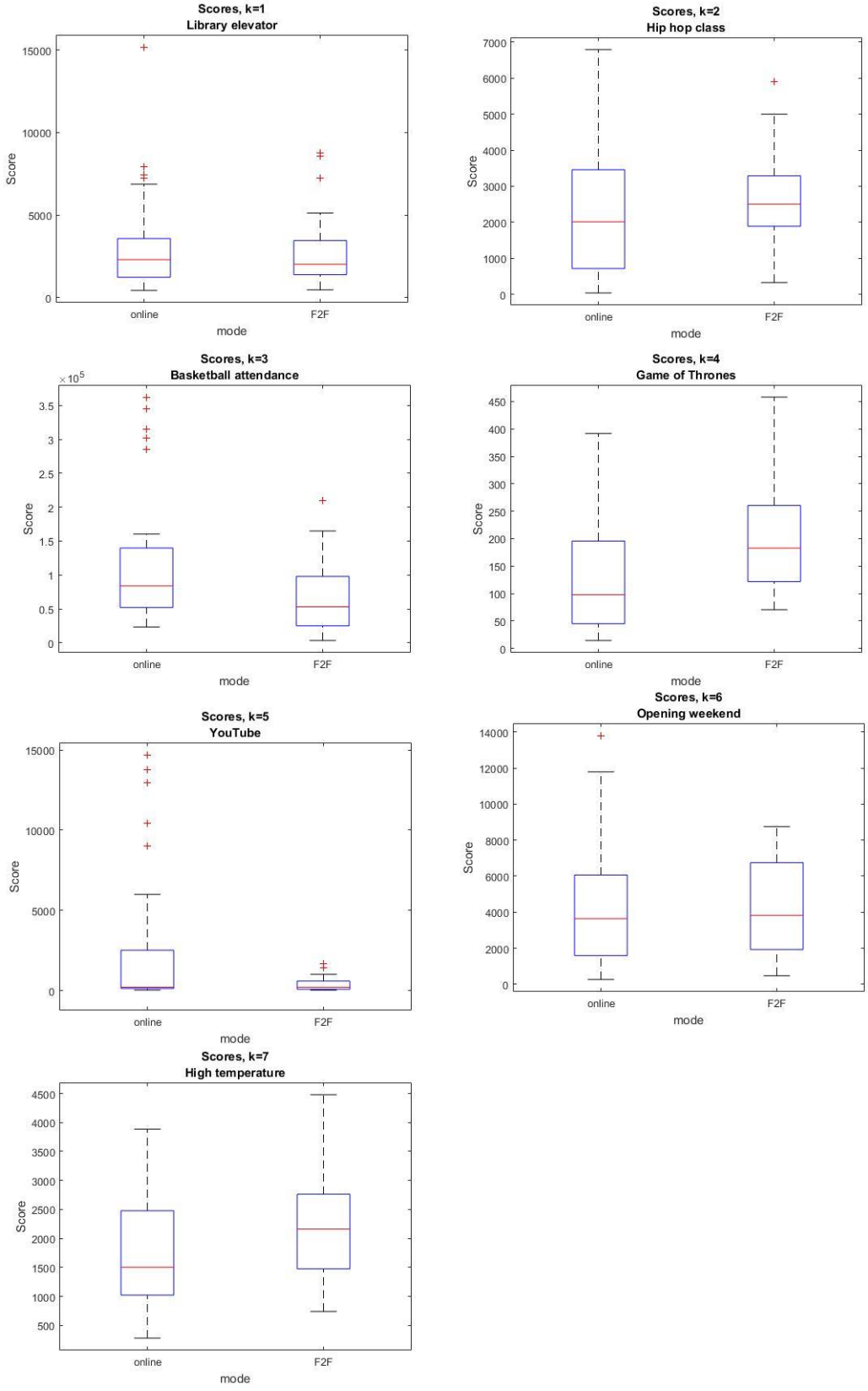
$$A_{ij2} = \left| x_2 - q_{a_1\, ij2} \right| (a_1) + \left| x_2 - q_{a_2\, ij2} \right| (a_2) + \left| x_2 - q_{a_3\, ij2} \right| (100 - a_3)$$

$$= (|11 - 8| \times 5) + (|11 - 30| \times 50) + \big(|11 - 40| \times (100 - 95)\big)$$

$$= 1110$$

We assigned the expert a core score of 1110. Our elicitation involved multiple experts, and comparing the core score of other experts responding to this question (Figure 6), we conclude that 1110 was a low score and therefore fairly accurate.

**Figure 6.** Boxplots comparing the effect of elicitation mode on the accuracy of forecasts.



### 3.5 Comparison of the Accuracy of the Forecasts Results

We used a question-by-question approach to examine the effect of elicitation mode on the accuracy of forecasts. Each participant was assigned a numerical value to represent their level of accuracy in responding to each question. Full details of our analysis, including descriptive statistics and graphs regarding individuals' core scores are found in Appendix M. The boxplots in Figure 6 compare the effect of mode, question-by-question, on the core scores (also see Appendix N for scatter plots of core scores). Following on, we carried out seven independent two sample tests, one test per question. The results are summarized in the Table 3 below.

We found the face-to-face elicitation provided accurate forecasts in four out of seven questions. In terms of statistically significant differences, there was no significant difference in four cases at the 5% level: the face-to-face elicitation was more accurate in two cases and online in one. However, at the 6% level, each of the modes was more accurate on two occasions. Thus, these results do not strongly support the idea that face-to-face elicitation are superior.

**Table 3.** Which mode resulted in more accurate values? Results from the independent two sample $t$-tests comparing online mean core score with the face-to-face.

| $k$ | **Question** | **More accurate?** | **Significantly lower score at 5% level?** | **Online M(SD)** | **F2F M(SD)** | $t$ | $df$ | $p$ (one-sided) | **Normality assumption holds?** (g*) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Library elevator | F2F | No | 3244 (3098) | 2703 (2015) | 0.82 | 48 | .2075 | No |
| 2 | Hip hop class | Online | No | 2288 (1818) | 2659 (1287) | 0.94 | 50 | .1758 | F2F yes |
| 3 | Basketball attendance | F2F | Yes | 119188 (100011) | 66505 (49427) | 2.63 | 40 | .006003 | No |
| 4 | Games of Thrones | Online | Yes | 124 (97) | 198 (89) | 3.19 | 59 | .001121 | No |
| 5 | YouTube | F2F | Yes | 2653 (4598) | 359 (394) | 2.72 | 29 | .005382 | No |
| 6 | Opening weekend | F2F | No | 4232 (3382) | 4192 (2521) | 0.05 | 52 | .4791 | No |
| 7 | High temperature | Online | No (Yes at 6%) | 1799 (1002) | 2191 (917) | 1.64 | 59 | .05285 | Yes $(-0.40)$ |

*g denotes Hedges measure of effect size with $|g| < 0.2$ "negligible", $|g| < 0.5$ "small"

It is important to note that the cores scores were measured on different scales and we cannot directly compare the core score across questions. For example, from the data in Table 3, at a glance it appears that the mean core scores from the Game of Thrones question are considerably lower, compared to the other scores. However, this does not suggest that on average the Game of Thrones forecasts produced the most accurate forecasts in this group of questions. The Game of Thrones forecasts were expressed in terms of "millions of views", however if we used "thousands of viewers" instead, the core scores would change considerably as shown in the Table 4.

**Table 4**. Core scores are scale-dependent. Example using the Game of Thrones question to illustrate the scale-dependent property of the multiple quantile scoring rule (Jose & Winkler, 2009).

| | Scale-dependent | |
|---|---|---|
| | Core score expressed in "millions of viewers" | Core score expressed in "thousands of viewers" |
| Online | 121 | 120827 |
| F2F | 199 | 198591 |

In order to combine the experts' scores and make a valid comparison, we required a scale-independent measure. A scale independent measure returns the same accuracy score, regardless if the forecasts were expressed in different units (for example "millions of viewers" or "thousands of viewers") (Jose, 2017). We achieved a scale-independent measure by transforming the experts' scores to a common scale using normalization.

### 3.6 Normalizing Scores Results

In this section we examine two linear, scale- independent transformations: min-max transformation and transforming the core scores by dividing by the mean. We compare the mean normalized score across questions.

Next, we explore how the transformation effects the shape of the underlying distribution. To do this, we consider the effect of the different transformations on two question separately: the library elevator and the high temperature question. The library elevator question is chosen because the underlying distribution is positively skewed and non-normal, whereas the high temperature question is chosen because the underlying distribution is normal.

The scatter plots in Figure 7 show the effect of the linear transformation. Both transformations have a minimum score of zero, in other words the transformation is anchored at zero (Osborne, 2002). The main difference between the linear transformations is the range: the min-max transformation was bounded between zero and one, however transforming the data by dividing by the mean has no upper bound.

Following on, we examine the shape of the underlying distribution using scatter plots (Figure 8) and we compare the mean, standard deviations, skew and kurtosis (Table 5). We found the linear transformations changed the mean and standard deviations, however had no effect on the skew and kurtosis (Osborne, 2002).

**Table 5.** Comparison of the different linear transformations.
(a) Accuracy cores from the library elevator question.

| | **Core scores** | **Normalized scores Divide by mean** | **Normalized scores Min max** |
|---|---|---|---|
| **Online** | | | |
| M (SD) | 4077 (4654) | 0.16 (0.21) | 1.19 (1.36) |
| Skew (kurtosis) | 2.46 (6.48) | 2.46 (6.48) | 2.46 (6.48) |
| **F2F** | | | |
| M (SD) | 2825(2121) | 0.11 (0.09) | 0.83 (0.62) |
| Skew (kurtosis) | 1.44 (1.34) | 1.44(1.34) | 1.44(1.34) |

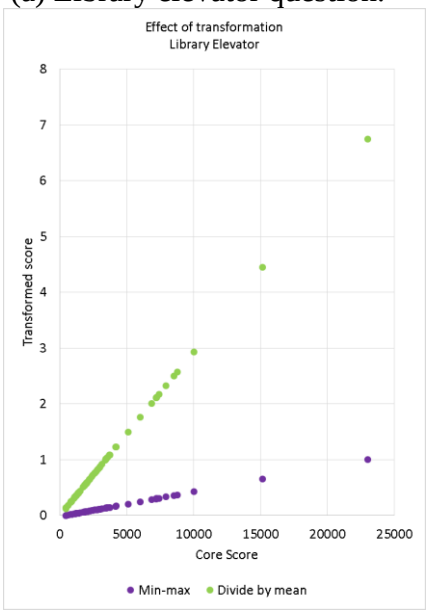
(b) Accuracy scores from the high temperature question.

| | **Core scores** | **Normalized scores Divide by mean** | **Normalized scores Min max** |
|---|---|---|---|
| **Online** | | | |
| M (SD) | 1834 (979) | 0.37 (0.23) | 0.91(0.48) |
| Skew (kurtosis) | 0.35 (-1) | 0.35(-1) | 0.35(-1) |
| **F2F** | | | |
| M (SD) | 2181 (928) | 0.45 (0.22) | 1.08 (0.46) |
| Skew (kurtosis) | 0.38 (-0.66) | 0.38 (-0.66) | 0.38 (-0.66) |

Figure 8 shows that when the min-max transformation was used then very different questions were still on a similar scale. In contrast, we divided by mean, then the library question had considerably higher scores. We argue that the min-max transformation is the valid approach for two reasons: the min-max transformation is a scale-independent measure and is bounded.

**Figure 7.** The effect of the linear transformations.

(a) Library elevator question.




(b) High temperature question.

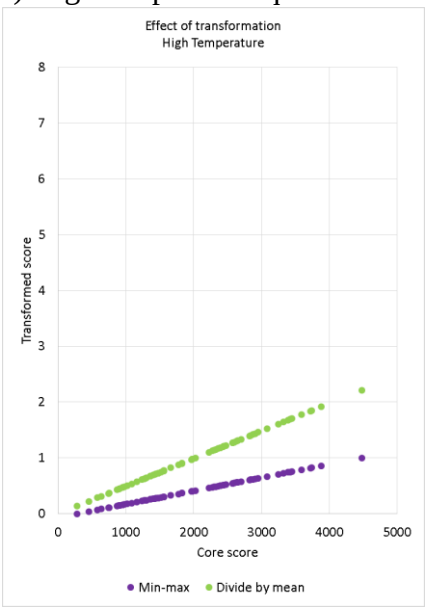
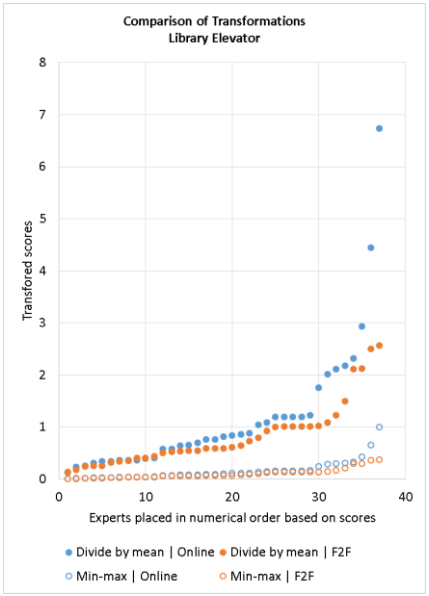



**Figure 8.** Scatter plots comparing the shape of the underlying distribution.

(a) Library elevator question.




(b) High temperature question.

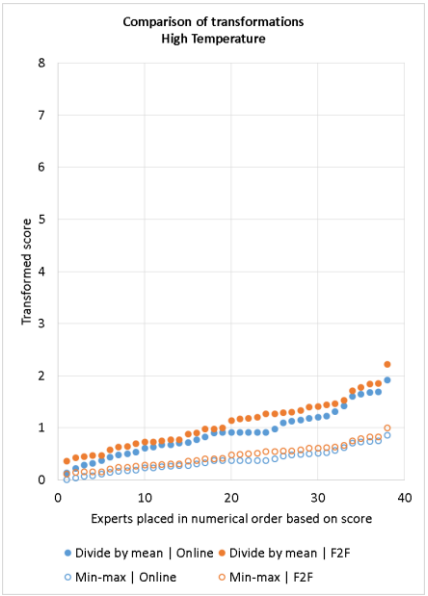
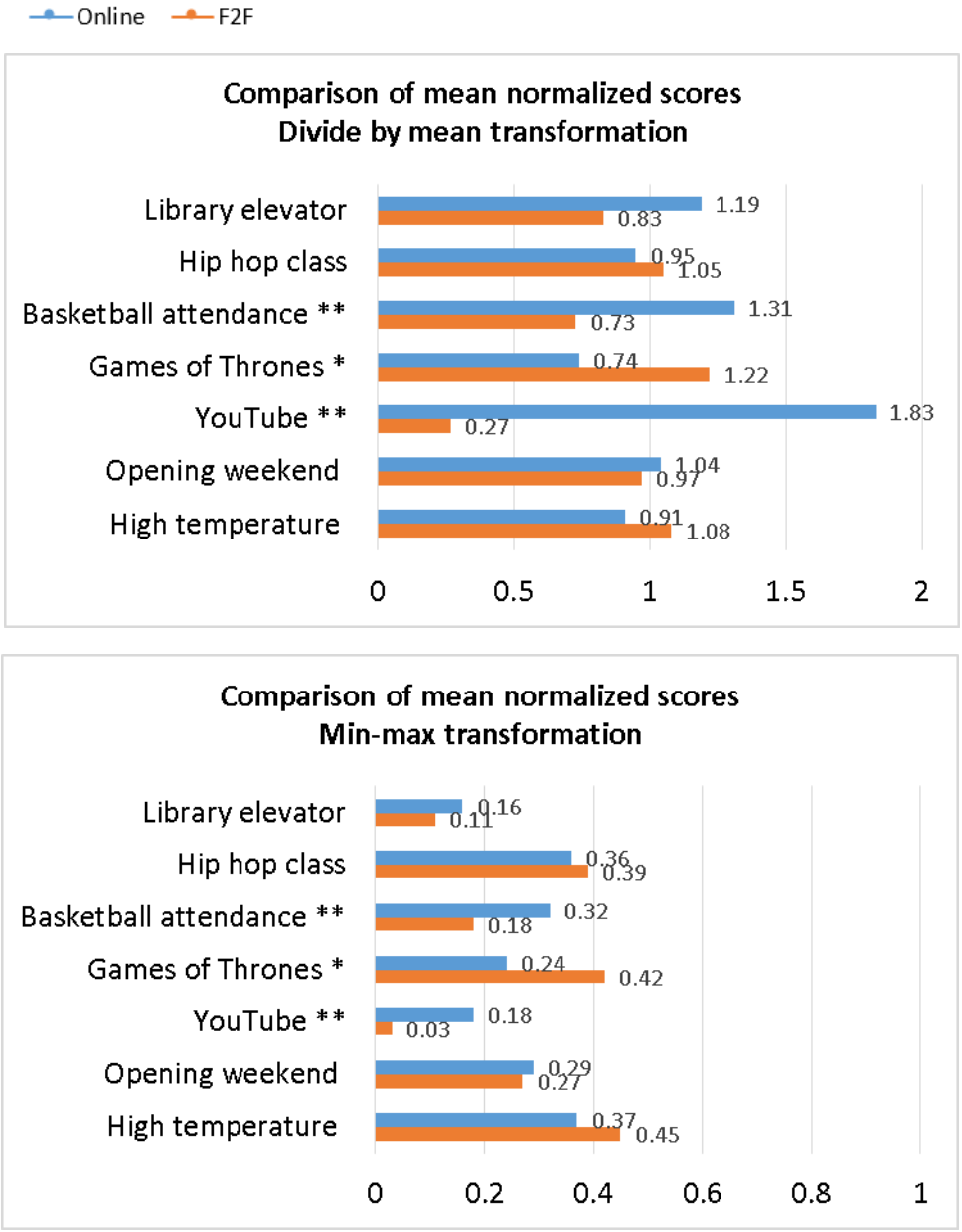


**Figure 9.** Comparison of mean normalized scores.




* The online elicitation obtained more accurate forecasts.
** The face-face elicitation obtained more accurate forecasts.

### 3.7 Comparison of the Accuracy of the Forecasts across Questions Results

In this section we present the results of our analysis when comparing the accuracy scores across questions using the min-max transformation. The min-max transformation was used for two reasons. First, it is scale-independent and second, it is bounded and therefore provided a well-defined range to compare the accuracy of the different questions (Tayman & Swanson, 1999).

We consider the scores, normalized using the min-max approach. While not statistically significant, we found that the face-to-face mean normalized score was lower than the online. Before we carried out an independent two sample $t$-test, we cleaned the data (Appendix F) and tested the assumption of normality.

In the case of the online normalized scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers (Figure 10). Review of the Shapiro-Wilk test for normality ($SW = 0.87, p < .001$), as well as the skewness $(1.11)$and kurtosis $(0.36)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal (Figure 10).

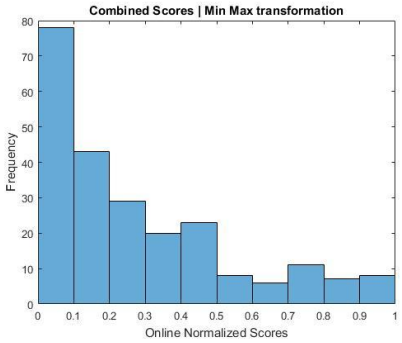
Regarding the face-to-face normalized scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers (Figure 10). Review of the Shapiro-Wilk test ($SW = 0.92, p < .001$), as well as the skewness $(0.8)$ and kurtosis $(0.03)$ statistics suggested the underlying distribution was not normal. The Q-Q

plot confirmed the result and we concluded the underlying distribution was non-normal (Figure 10).

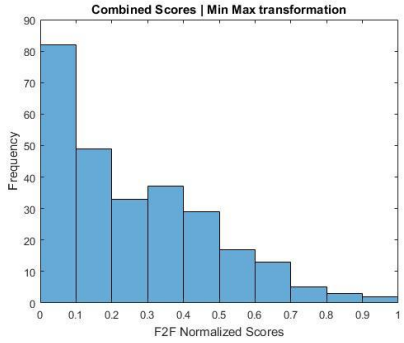
Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(453) = 0.51, p = .3034$ (two-sided). The result indicated the mean online normalized scores ($M = 0.27, SD = 0.26\,, n = 233$) was not significantly higher than the mean face-to-face elicited normalized scores ($M = 0.26, SD = 0.22\,, n = 270$).

**Figure 10.** Comparing the effect of mode on the accuracy of forecasts across questions. Core scores are normalized using the min-max transformation.

(a) Online normalized scores.

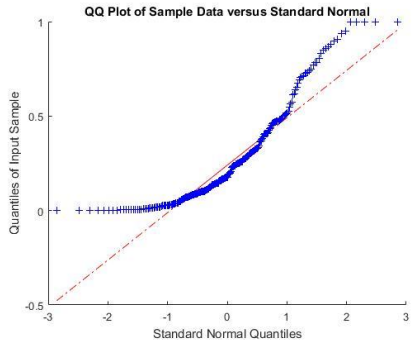



(b) Face-to-face normalized scores.

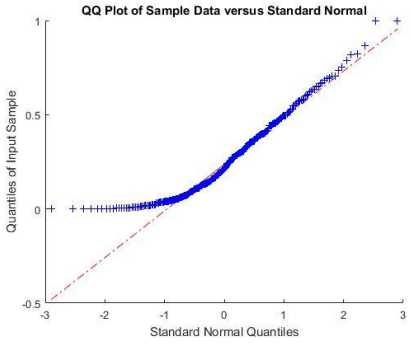



(c). Box plots.

(d) Q-Q plot. Online normalized scores




(e) Q-Q plot. F2F normalized scores.




### 3.8 Comparison of the Presence of Satisficing Results

In this section we investigate the effect of satisficing towards the end of the elicitation (late) compared to at the beginning of the elicitation (early). We used the "library elevator" question ($k$ = 1) to make the comparison because the question appeared near the beginning of the elicitation (question 3) and near the end (question 18). Participants were asked to make a judgement on the time taken to travel in the elevator to the 23rd floor of the library. The face-to-face elicitation performed as expected, with the more accurate forecasts obtained from questions appearing early in the elicitation; we note however, that there was little difference between the two face-to-face groups (early compared to late). On the other hand, we found the opposite result from the online elicitation: results from later in the elicitation were better. The Figure 11 below shows the aggregated probability distributions for the library elevator question. In Figure 11 we observe a more pronounced difference in the online elicitation between the elicited percentiles obtained early in the process compared to late. Examining the core scores from the online elicitation, the boxplot in Figure 12 shows that while not statistically significant, the higher mean core score was found when the question appeared early in the elicitation. In

other words, the more accurate forecasts were obtained when the question is at the end, as opposed to at the beginning of the online elicitation. It seems to suggest that participants improved as they moved through the online tool.

**Table 6.** Descriptive statistics of the core scores from the library elevator question.

| **Variable** | $\boldsymbol{n}$ | $\boldsymbol{M}$ | $\boldsymbol{SD}$ | **Min** | **Max** |
|---|---|---|---|---|---|
| Online, early | 14 | 4926 | 6390 | 795 | 23020 |
| Online, late | 19 | 3451 | 2837 | 427 | 10005 |
| F2F, early | 20 | 2775 | 2210 | 465 | 8765 |
| F2F, late | 17 | 2884 | 2077 | 615 | 7250 |

The scatter plot of the core scores for the library elevator question (Figure 14), compares the accuracy of questions appearing early compared with late. The sample size was adjusted to compare samples of equal size. In the case where the sample size was not equal to 20, the mean score was added.

An independent two sample $t$-test was conducted to compare the mean core scores of a question appearing early in the elicitation, with the same question appearing. First we tested the online core scores to find out if early presentation obtained in a significantly higher core score, and hence less accurate forecast compared than late. Results indicated the mean core score from when the question appeared early in the online elicitation was not significantly higher than ($M = 4926,\ SD = 6390,\ n = 14$) when the question appeared late ($M = 3451,\ SD = 2837,\ n = 19$), $t(16) = 0.81, p = .2155$ (one-sided).

Next, we tested the face-to-face core scores. In this case we tested to find out if early presentation obtained a lower core score on average, and hence more accurate forecast.

Results indicated the mean scores from when the question appeared early in the face-to-face elicitation was not significantly lower than ($M = 2775, SD = 2210, n = 20$) when the question appeared late ($M = 2884, SD = 2077, n = 17$), $t(34) = 0.15, p = .4389$ (one-sided).

**Figure 11.** Aggregated probability distribution for the library elevator question. The observed value was 63 seconds.

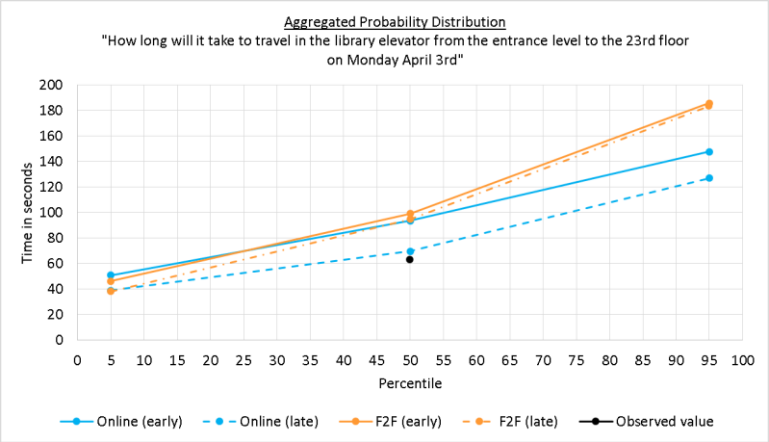



**Figure 12.** Boxplots comparing the core scores from the library elevator question.

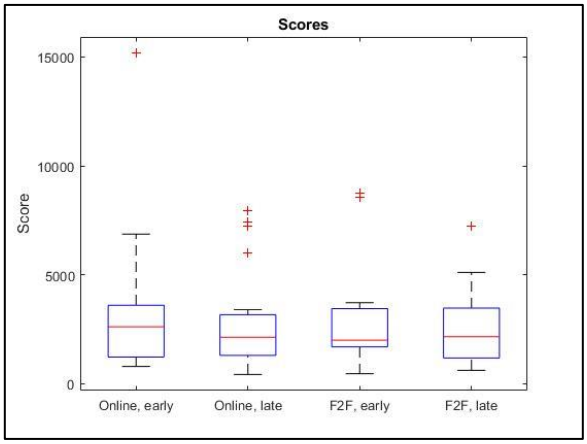


**Figure 13.** Scatter plot of core scores for library elevator question.

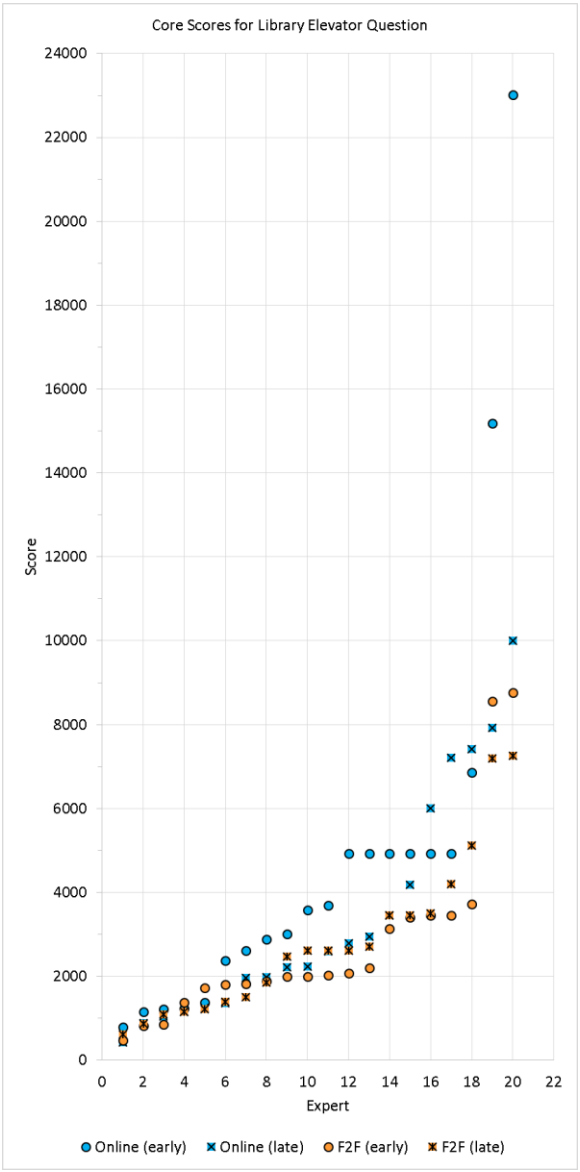


**Table 7.** Questions appearing early in the elicitation obtained more accurate values?

(a) Core Scores library elevator question.

| | | | **Core Scores Library Elevator question** | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Mode** | **More accurate?** | **Significant difference?** | $\boldsymbol{M}$ $(\boldsymbol{SD})(\boldsymbol{n})$ **Early** | $\boldsymbol{M}$ $(\boldsymbol{SD})(\boldsymbol{n})$ **Late** | $\boldsymbol{t}$ | $\boldsymbol{df}$ | $\boldsymbol{p}$ **(one-sided)** | **Normality assumption holds?** |
| Online | Late | No | 4926 (6390) (14) | 3451 (2837) (19) | 0.81 | 16 | .2155 | No |
| F2F | Early | No | 2775 (2210) (20) | 2884 (2077) (17) | 0.15 | 34 | .4389 | No |

(b) Core Scores library elevator question with 10% trim

| | | | **Core Scores Library Elevator question (10% trim)** | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Mode** | **More accurate?** | **Significant difference?** | $\boldsymbol{M}$ $(\boldsymbol{SD})(\boldsymbol{n})$ **Early** | $\boldsymbol{M}$ $(\boldsymbol{SD})(\boldsymbol{n})$ **Late** | $\boldsymbol{t}$ | $\boldsymbol{df}$ | $\boldsymbol{p}$ **(one-sided)** | **Normality assumption holds?** (g*) |
| Online | Late | No | 3762 (3932) (12) | 3244 (2398) (17) | 0.41 | 16 | .3448 | No |
| F2F | Early | No | 2206 (850) (16) | 2744 (1783) (15) | 0.86 | 19 | .801 | Yes $(-0.3)$ |

* g denotes Hedges measure of effect size with $|g| < 0.2$ "negligible", $|g| < 0.5$ "small"

### 3.9 Summary of Results

In this section we summarize our results. Overall, we found no statistically significant difference in the means of the median estimates; the uncertainty range; the level of overconfidence; the accuracy of the forecasts and the presence of satisficing. Table 8 summarizes our findings.

**Table 8.** Summary of the actual findings. The research questions from the original proposal are re-printed here (Baker, 2016, p7).

| **Research Question** | **Values or metrics used** | **Hypothesis** | **Actual findings** |
|---|---|---|---|
| Did different modes led to different central values? | Means of median estimates. | No difference. | No significant difference. |
| Which mode resulted in a larger uncertainty range and less overconfidence? | Uncertainty range and overconfidence. | F2F will have a larger uncertainty range and less overconfidence. | No significant difference. |
| Which mode resulted in more accurate values? | Multiple quantile scoring rule (Jose & Winkler, 2009). | F2F will have more accurate results. | No significant difference. |
| Which mode produced satisficing? | Multiple quantile scoring rule (Jose & Winkler, 2009). | F2F will have less satisficing. | No significant difference. |

# CHAPTER 4

# CONCLUSION

This research project compared two elicitation modes: the traditional face-to-face elicitation interview against an equivalent online elicitation survey. Differences in central values, overconfidence, accuracy and satisficing were measured. In this section we summarize our findings.

First, we considered if the use of different elicitation modes generated different central values. We analyzed the mean median estimates, question by question, and found no statistically significant difference between modes in five out of seven questions. The evidence is not sufficient to conclude that the different modes, in any systematic way, obtained different central values.

Second, we investigated if a particular elicitation mode would be more effective at limiting overconfidence. The differences between modes comparing the uncertainty range and then the rate of surprises were not statistically significant. Specifically, while the average uncertainty range from the online distributions was slightly higher, indicating the online forecasts expressed less overconfidence, it was not significantly higher. On the other hand, while the face-to-face elicitation obtained a lower rate of surprises, this time implying the face-to-face forecasts expressed less overconfidence, the rate was not significantly lower. In sum, we found no statistically significant difference in the level of overconfidence and the evidence here does not show an advantage to either mode.

Beyond analyzing the tails of the elicited distributions, we quantified the accuracy of each mode by combining the observed value and the three elicited quantile assessments. We referred to this quantity as the core score. We used a question by question approach, based on the core score, to compare the modes. We found that the modes are nearly evenly split. In other words, we found the face-to-face elicitation mode to be significantly more accurate in two instances and the online in two at the 6% level. In addition, we aggregated the normalized core scores across questions and found no statistically significant difference in the mean normalized core scores. Thus, again we found no evidence to show that the face-to-face elicitation mode was superior.

Finally, we hypothesized that the online elicitation mode would lead to more satisficing. In order to detect the presence of satisficing, we used the core scores and compared response order. We compared the average core score of a question presented early in the elicitation against late. Surprisingly, we found that the online forecasts obtained late in the elicitation were more accurate. Both modes, however, showed no statistically significant difference in average cores scores between early and late response order.

In short, the results of our analysis indicate the differences between the accuracy of an online elicitation and face-to-face elicitation are not significant, and consequently the online elicitation mode may be used successfully to obtain accurate forecasts. However, a limitation of our analysis is that it focuses on a subset of the elicitation questions. Although, we spent time adapting the face-to-face protocol to an equivalent online elicitation, we found on completion of the elicitations, that only 35% of the questions

were directly comparable. This highlights that the online elicitation mode needs further improvement to be able to respond more effectively in real-time. Our design of the online elicitation meant decisions relating to the most appropriate units of measurement (for example minutes or seconds) as well as the bounds of the unknown parameter, were fixed beforehand. Future work might concentrate on improvements to the online elicitation design. In particular, research might investigate strategies to limit the anchoring and adjustment heuristic, as well as the impact of enhancing the interactive ability of the software and allowing individual users flexibility to personalize the elicitation settings, for example to select preferred units of measurement.

## APPENDIX A

## ELICITATION QUESTIONS

**Table 9.** Elicitation questions

| Theme | Question Title | Question |
|---|---|---|
| UMass library | Library reserve desk | How long will a person wait in line at the Circulation/Reserve desk on the lower level of the Du Bois Library on Monday April 3rd? |
| | Library computer | How long will a person wait to use a computer in the public workstation of the Du Bois Library at 12:45pm on Monday April 3rd? |
| | Library elevator | How long will it take to travel in the library elevator from the entrance level to the 23rd floor on Monday April 3rd? |
| UMass required courses | GenEd grades | What is the grade of a randomly chosen student taking a class that satisfies their Social & Behavioral General Education requirement in the spring semester? |
| | Enrolled Chemistry | How many students will be enrolled in Chemistry 111 on Monday February 6th 2017, the last day students can drop class for the spring semester? |
| | English grades | How many undergraduate students will receive a grade of C+ or below in the College Writing course, English 112, at UMass Amherst for spring semester 2017? |
| UMass recreation center | Jogging track | How many people will use the Jogging/ Walking Track at the Recreation Center on Monday April 3rd at 7:00pm? |
| | Hip hop class | How many people will participate in the Cardio Hip Hop class at 7:30pm on Monday 3rd April? |
| | Rec center | How many people will use the Recreation Center on Monday April 3rd between 6pm and 6:30pm? |
| UMass men's basketball | Basketball attendance | What will be the recorded attendance at the UMass Men's Basketball vs. Richmond on March 1st, 2017? |
| | Basketball game | How long will it take for all patrons to leave the seating area marked section T in the Mullins center after the UMass Men's Basketball vs. Richmond on March 1st 2017? |
| Entertainment | Game of Thrones | How many people will tune in for the Season 7 premiere of HBO's Game of Thrones telecast in spring of 2017? |

| | | |
|---|---|---|
| | Tweets | How many tweets will be made per second on Friday March 31st, from 10:30am to 10:35am (Eastern Time)? |
| | YouTube | How many YouTube videos will be viewed on Friday March 31st, from 11am to 11:05am (Eastern Time)? |
| | Opening weekend | How much will the movie Guardians of the Galaxy Vol. 2, due to be released on May 5th, earn in the U.S. over its opening weekend? |
| UMass dining | Stir fry | How long will a person wait to be served stir fry at the International Cuisine station in the Berkshire dining hall on Monday April 10th? |
| | High temperature | What will the high temperature be in Amherst as measured at Amherst College on April 10th, 2017? |
| | Ice cream cold day | How long will the wait be for ice cream in the Blue Wall Cafe on Monday April 10th? Assume that the high temperature that day in Amherst is below 70°F. The wait time will be measured from the first person arriving between 12.50pm and 1.10pm on Monday April 10th, to when they are served. |
| | Ice cream warm day | How long will the wait be for ice cream in the Blue Wall Cafe on Monday April 10th? Assume that the high temperature that day in Amherst is above 70°F. |
| | Pizza delivery | How long will it take for a medium pepperoni pizza to be delivered to Marston Hall from Bruno's Pizza restaurant on Thursday April 13th, 2017 at 6pm? The delivery time will be measured from the end of the phone order, to handing over the pizza box. |

## APPENDIX B

## ELICITATION QUESTION ORDER

**Table 10.** Elicitation question order.

| **Question Order 1a F2F** | | **Question Order 1b Online** | | **Question Order 2 F2F & Online** | |
|---|---|---|---|---|---|
| 1 | Library reserve desk | 1 | Library computer | 1 | Pizza delivery |
| 2 | Library computer | 2 | Library reserve desk | 2 | High temperature |
| 3 | Library elevator | 3 | Library elevator | 3 | Ice cream cold day |
| 4 | GenEd grades | 4 | Enrolled Chemistry | 4 | Ice cream warm day |
| 5 | Enrolled Chemistry | 5 | English grades | 5 | Stir fry |
| 6 | English grades | 6 | GenEd grades | 6 | Game of Thrones |
| 7 | Jogging track | 7 | Rec center | 7 | YouTube |
| 8 | Hip hop class | 8 | Jogging track | 8 | Opening weekend |
| 9 | Rec center | 9 | Hip hop class | 9 | Tweets |
| 10 | Basketball attendance | 10 | Basketball game | 10 | Basketball attendance |
| 11 | Basketball game | 11 | Basketball attendance | 11 | Basketball game |
| 12 | Game of Thrones | 12 | Opening weekend | 12 | Jogging track |
| 13 | Tweets | 13 | Game of Thrones | 13 | Hip hop class |
| 14 | YouTube | 14 | YouTube | 14 | Rec center |
| 15 | Opening weekend | 15 | Tweets | 15 | Enrolled Chemistry |
| 16 | Stir fry | 16 | High temperature | 16 | GenEd grades |
| 17 | High temperature | 17 | Ice cream cold day | 17 | English grades |
| 18 | Ice cream cold day | 18 | Ice cream warm day | 18 | Library elevator |
| 19 | Ice cream warm day | 19 | Stir fry | 19 | Library reserve desk |
| 20 | Pizza delivery | 20 | Pizza delivery | 20 | Library computer |

## APPENDIX C

## EXAMPLE OF INTERVIEW SCRIPT FROM F2F ELICITATION (QUANTILE ASSESSMENT)

Below is an excerpt from the face-to-face interview script. In this case, quantile assessment was used to gather the subjective probability distributions. Instructions are in italics.

**Library elevator**

How long will it take to travel in the library elevator from the entrance level to the 23rd floor on Monday April 3rd? The travel time will be measured from when the first up elevator's door closes, after 11:00am, to when the elevator doors open at the 23rd floor.

To give you some background information:

- Travel time was 27 seconds on Thursday January 19th at 8:00am.
- The travel time on Monday 23rd January was one minute, six seconds.

Do you think the travel time will be longer, shorter, or about the same on Monday April 3rd compared to Monday January 23rd? Why would it change?

a) What is the shortest travel time? We are looking for a value that is sufficiently small that you think there is perhaps only 1 chance in 20 that the actual travel time will turn out to be shorter. Why? Please provide a numerical answer and a rationale if possible.

(b) Now let's look at the other extreme, what is the longest travel time? Now we are looking for a value that is sufficiently large that you think there is perhaps only 1 chance in 20 that the actual travel time will be longer. Why? Please provide a numerical answer and a rationale if possible.

*(Part (a) & (b) gives us a range of possible outcomes.)*

You are saying that you think there is about a 90% chance that the travel time will be between [*insert the answer from part (a)]* and [*insert the answer from part (b)*].

*(If this is correct, go on. If you feel that you'd like to rethink (a) & (b), please go back.)*

(c) Now that we have a range you are comfortable with, let's talk about what your "break-even" bet would be. What is the travel time that you think is about the $50^{th}$ percentile?

That it is equally likely that the true travel time will be less than or greater than?

## APPENDIX D

## EXAMPLE OF INTERVIEW SCRIPT FROM F2F ELICITATION (PROBABILITY ASSESSMENT)

Below is an excerpt from the face-to-face interview script. In this case, probability assessment was used to gather the subjective probability distributions. Instructions are in italics.

**GenEd grades**

What is the grade of a randomly chosen student taking a class that satisfies their Social & Behavioral General Education requirement in the spring semester?

What is the probability that a randomly selected student will achieve a grade of A, A-?

What is the probability of a B+, B, B-?

What is the probability of a C+, C, C-?

What the probability of a grade of D or lower?

Okay, let's just check a couple of things.

You think that more than half of the students will receive a grade of [*insert the correct statement: "B- or better", or "C+ or worse"*]. Is that correct?

Also, you think the most students will get [*insert the highest probability*] and that the fewest number of students will get [*insert the lowest probability*]. Is that correct?

## APPENDIX E

## QUESTION GROUPS FOR ANALYSIS

**Table 11.** Question groups for analysis. Questions are organized into four groups for analysis.

(a) Question Group 1.

| **Question** | **Observed Value** | **Online range** |
|---|---|---|
| Library elevator ($k = 1$) | 63 seconds | 20 to 280 seconds |
| Hip hop class ($k = 2$) | 11 people | 5 to 70 people |
| Basketball attendance ($k = 3$) | 2434 people | 1000 to 11000 people |
| Game of Thrones ($k = 4$) | 10.11 million viewers* | 1 to 10 million viewers |
| YouTube ($k = 5$) | 20.6313 million videos | 10 to 140 million videos |
| Opening weekend ($k = 6$) | 146.51 million dollars | 20 to 240 million dollars |
| High temperature ($k = 7$) | 79 °F | 10 to 90 °F |

* The Game of Thrones question was included in group 1 although the observed value fell outside the online range. We included the question in our analysis because HBO allowed viewers to watch the episode free of charge and responses from the face-to-face elicitation (Appendix S) gave a reduction in the HBO subscription fee as a reason for the observed number of viewers turning out to be higher than their high estimate.

(b) Question Group 2. The observed value fell outside the range displayed in online elicitation.

| **Question** | **Observed value** | **Online range** |
|---|---|---|
| Jogging track | 10 people | 20 to 280 people |
| Rec center | 218 people | 1000 to 9000 people |
| Basketball game | 3.3 minutes | 20 to 220 minutes |
| Tweets | 7.57 thousand per second | 10 to 130 thousand per second |

(c) Question Group 3. The observed value was expressed in different units from stated on the online elicitation. For example the observed was expressed in terms of seconds however the online elicitation asked for responses to be expressed in minutes.

| Question | Observed value | Online range |
|---|---|---|
| Library reserve desk | 0.05 minutes | 10 to 80 minutes |
| Library computer | 2.683 minutes | 10 to 80 minutes |
| Stir fry | 3.73 minutes | 10 to 80 minutes |
| Ice cream warm day | 1.56 minutes | 10 to 80 minutes |

(d) Question Group 4.

| Question | Reason not included in analysis |
|---|---|
| Enrolled Chemistry | Error with online elicitation. |
| GenEd grades | No observed value. |
| English grades | Variation in wording. |
| Ice cream cold day | Temperature in Amherst was above 70°F. |
| Pizza delivery | Error with online elicitation. |

## APPENDIX F

## A LIST OF EXPERTS REMOVED FROM THE ANALYSIS

**Table 12.** A list of experts removed from the analysis. Experts were removed from our analysis because their subject probability distributions were incomplete.

| **Question** | **List of experts removed from analysis** | **Online** $n_{1k}$ | **F2F** $n_{2k}$ |
|---|---|---|---|
| Library elevator | 22, 57 & 59 | 33 | 37 |
| Hip hop class | | 34 | 39 |
| Basketball attendance | 16 | 33 | 39 |
| Game of Thrones | 17 | 33 | 39 |
| YouTube | | 34 | 39 |
| Opening weekend | | 34 | 39 |
| High temperature | 17, 33 and 69 | 32 | 38 |
| **Total number of forecasts in group 1:** | | $F_{11}$ =267 | $F_{21}$ =309 |

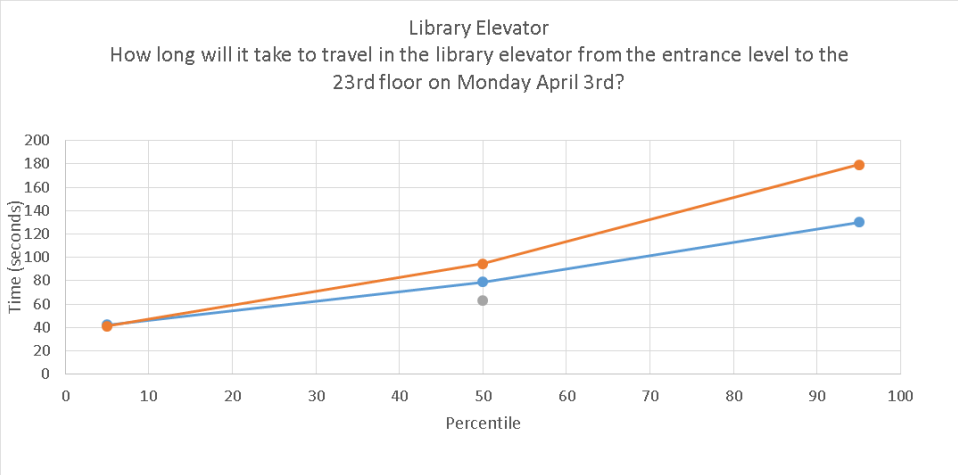
## APPENDIX G

## AGGREGATED PROBABILITY DISTRIBUTIONS

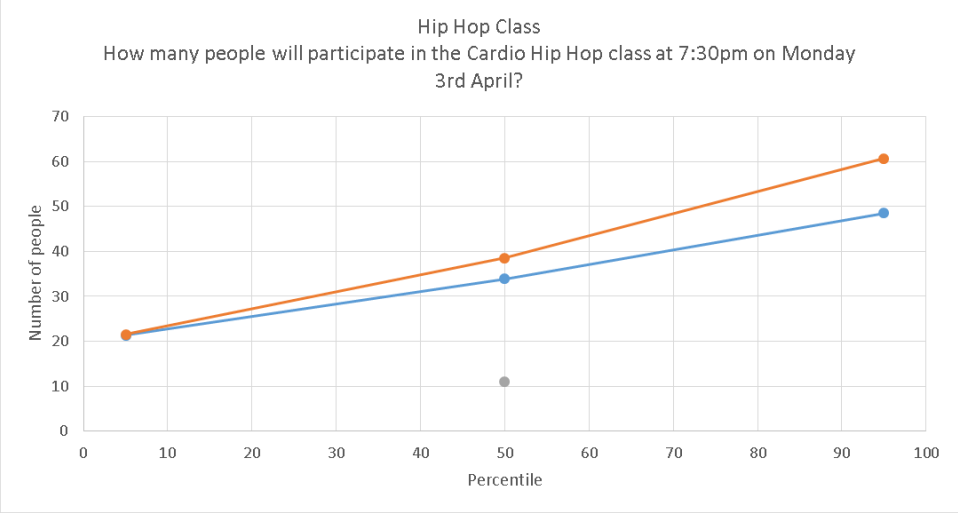
**Figure 14.** The aggregated probability distributions for each question in group 1.

Key Online F2F Observed value

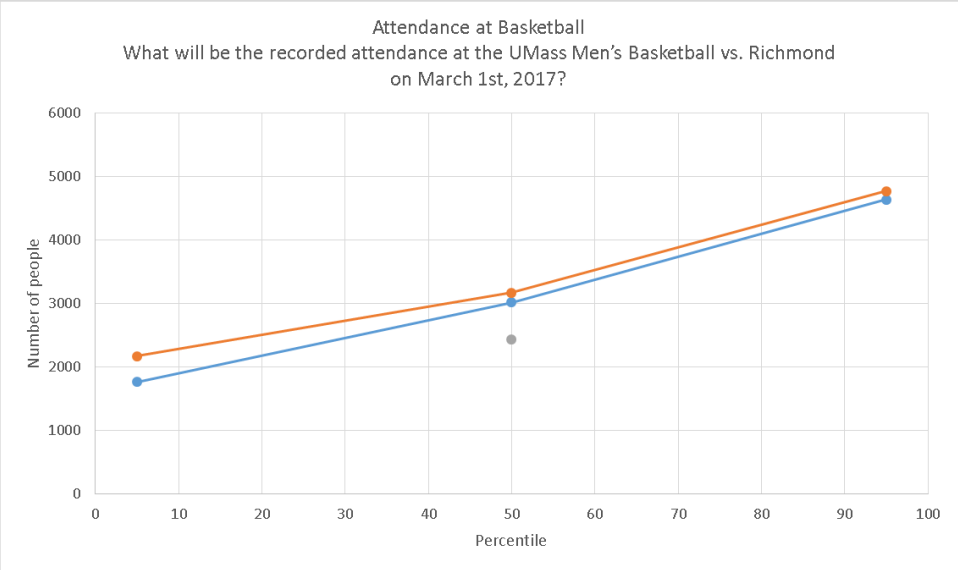
(a) The aggregated probability distribution for the Library Elevator question.



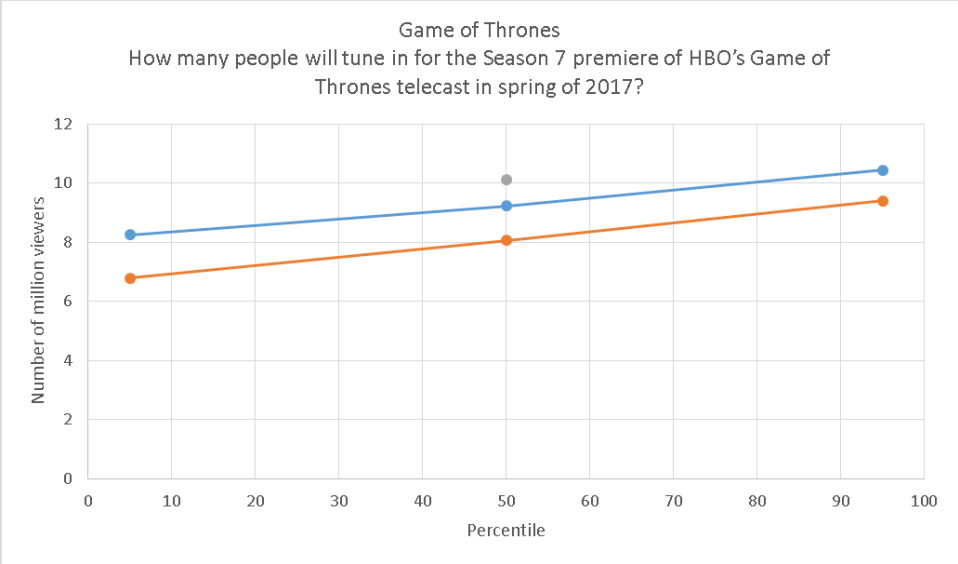

(b) The aggregated probability distribution for the Hip Hop Class question.



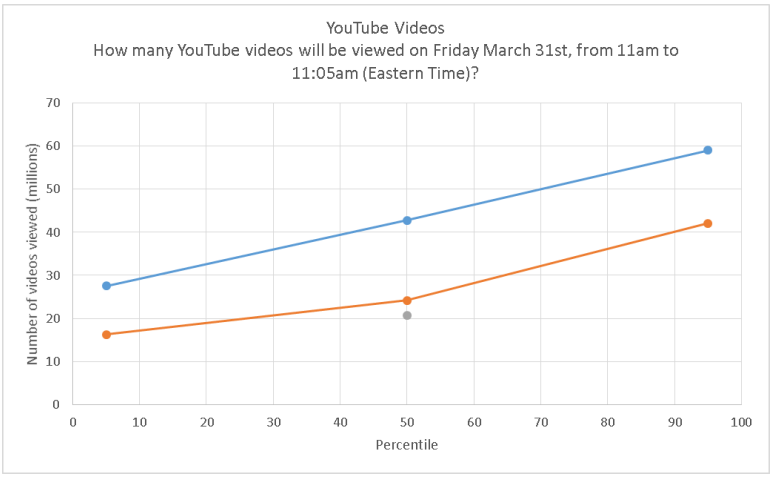
(c) The aggregated probability distribution for the Basketball Attendance question.



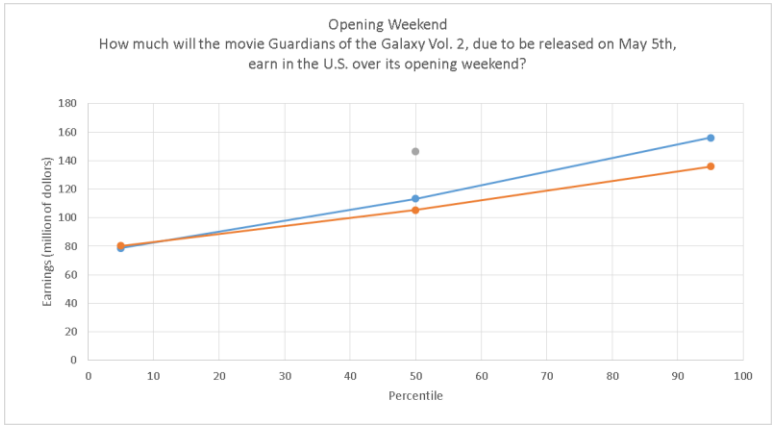

(d) The aggregated probability distribution for the Game of Thrones Question.



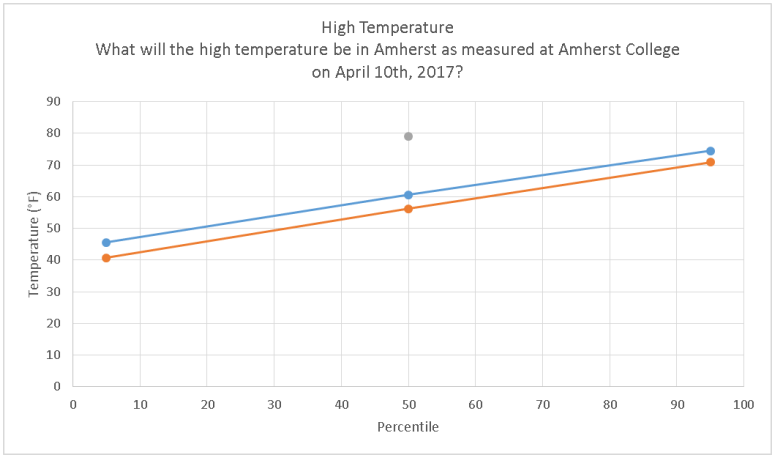
(e) The aggregated probability distribution for the YouTube question.




(f) The aggregated probability distribution for the Opening Weekend question.



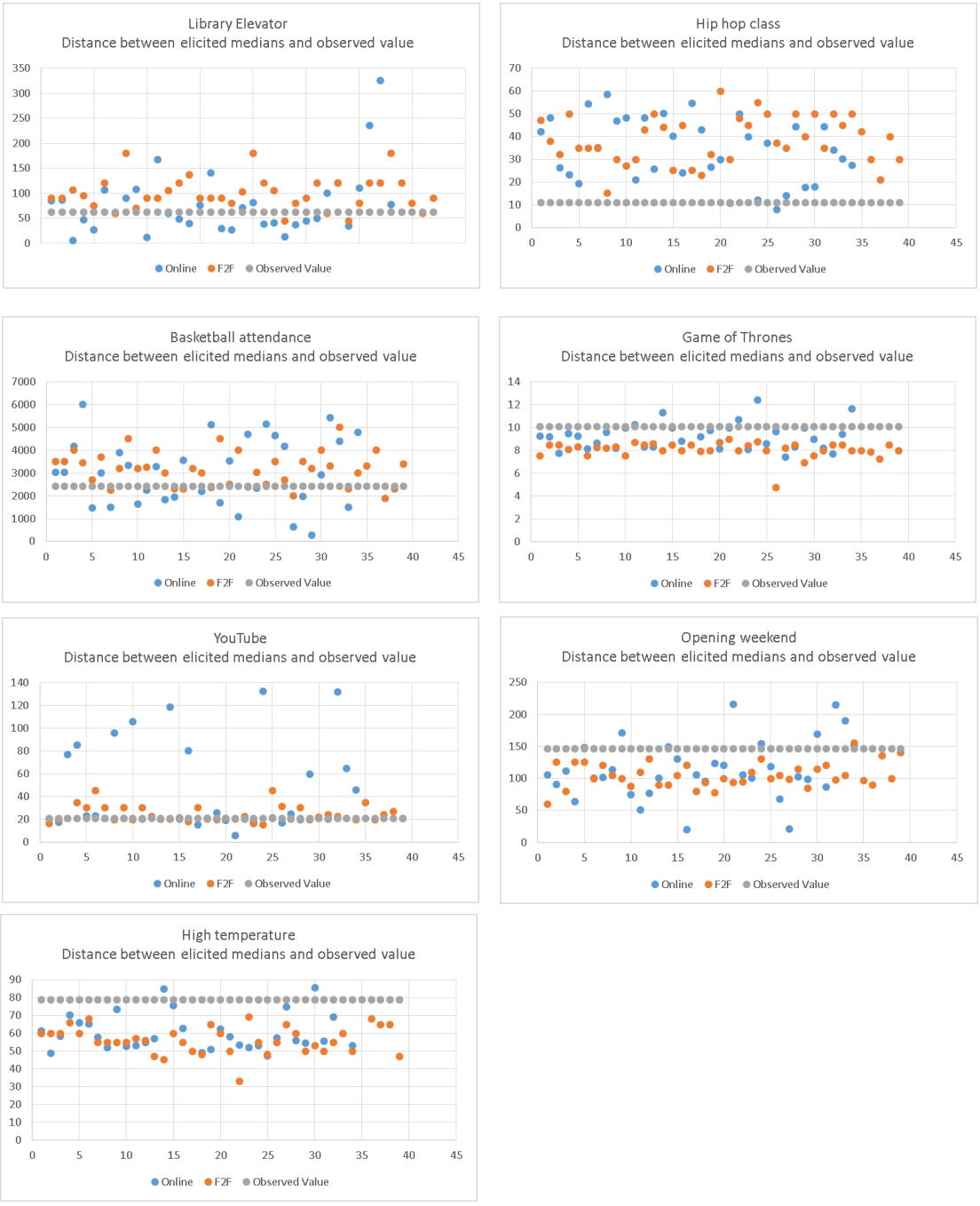

(g) The aggregated probability distribution for the High Temperature question.



# APPENDIX H

## MEDIAN ESTIMATES

**Figure 15.** The vertical distance between the median estimate and the observed value.



# APPENDIX I

## COMPARISON OF THE CENTRAL VALUES

The analysis of the difference between the central values was approached question by question.

### I.1 Comparison of the central values gathered from the library elevator question.

We considered the elicited median values from the library elevator question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 22, 57 and 59 were removed from the analysis because their probability distributions were incomplete.

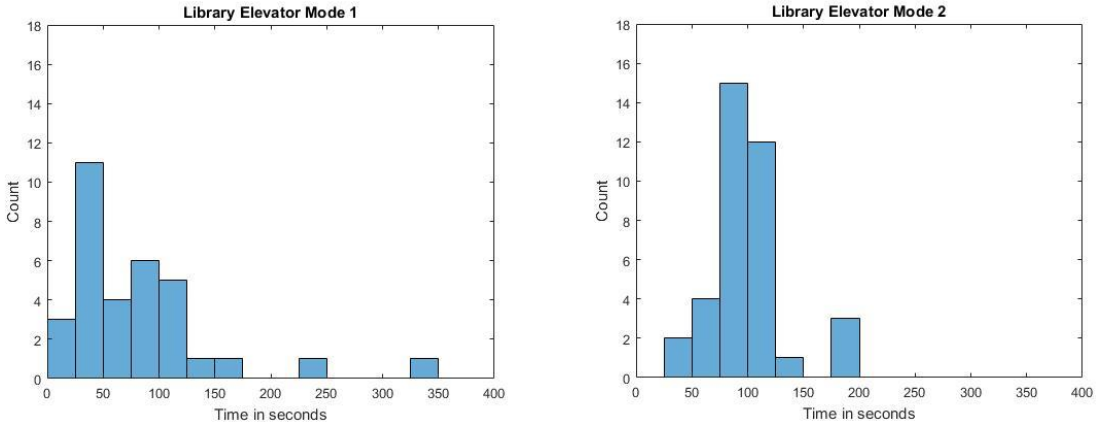
In the case of the online elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.80, p < .001$), as well as the skewness $(1.96)$ and kurtosis $(4.44)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.91, p = .005903$), as well as the skewness $(0.8)$ and kurtosis $(0.59)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

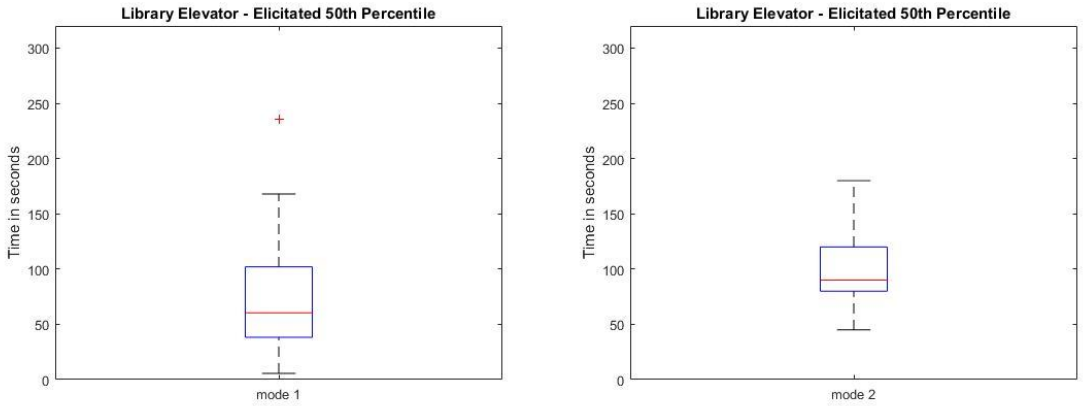
Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(45) = -1.67, p = .101$ (two-sided). This result indicates no significant difference between the online mean elicited median values ($M = 79, SD = 66, n = 33$) and the face-to-face elicited median values ($M = 100, SD = 33, n = 37$).

**Figure 16.** Comparison of the central values gathered from the library elevator question.

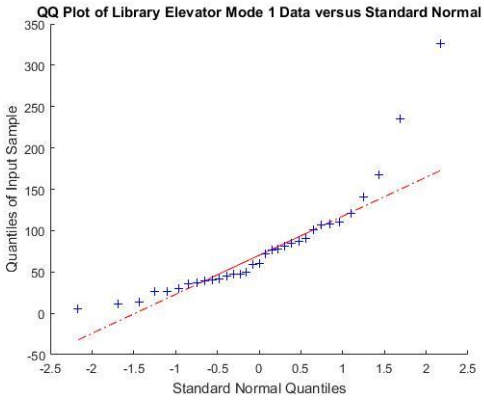
(a) Histograms. Elicited 50th percentile (median) from the library elevator question.

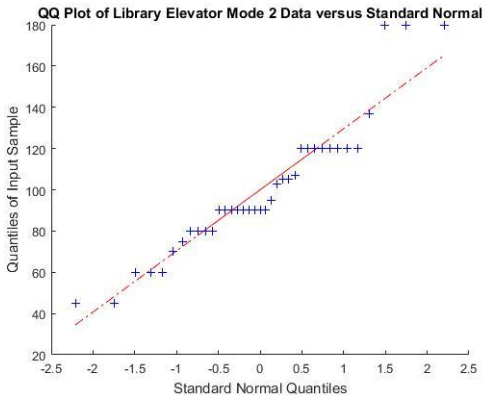



(b) Boxplots. Elicited 50th percentile (median) from the library elevator question.



(c) Quantile-quantile plots (Q-Q Plots) indicate the normality assumption does not hold.




**I.2 Comparison of the central values gathered from the hip hop class question.**

We considered the elicited median values from the hip hop class question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

In the case of the online elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality($SW = 0.96 , p = .2544$ ), as well as the skewness $(-0.1)$and kurtosis $(-1.24)$ statistics suggested the underlying distribution is normal. The Q-Q plot confirmed the result and we concluded that the underlying distribution was normal.

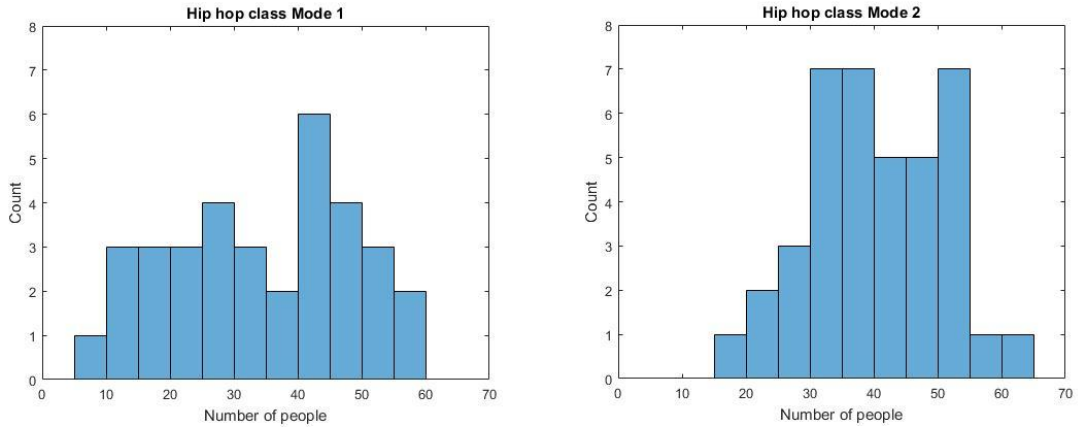
Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.97 , p = .4674$), as well as the skewness $(-0.12)$ and

kurtosis $(-0.81)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we conclude the underlying distribution was normal.

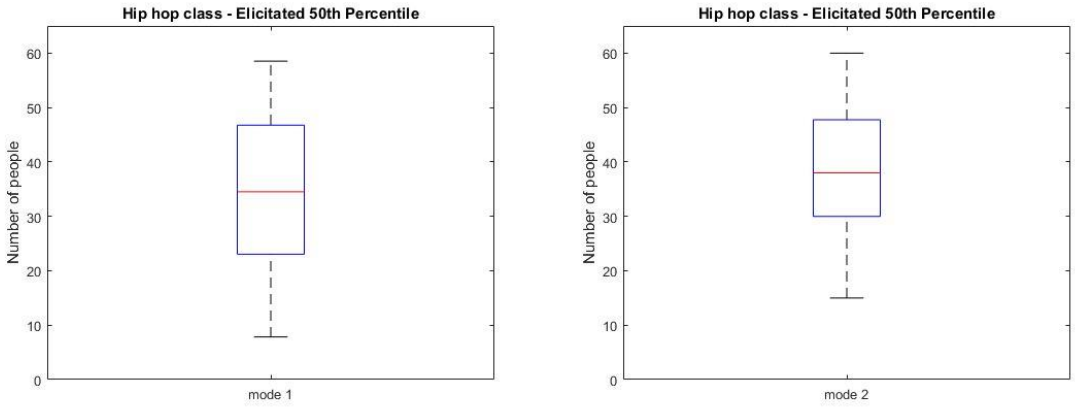
The underlying distributions were normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(59) = -1.59, p = .1181$ (two-sided); $g = -0.38$ (small). This result indicated no significant difference between the online mean elicited median values $(M = 34, SD = 14, n = 34)$ and the face-to-face mean elicited median values $(M = 39, SD = 10, n = 39)$.

**Figure 17.** Comparison of the central values gathered from the hip hop class question.

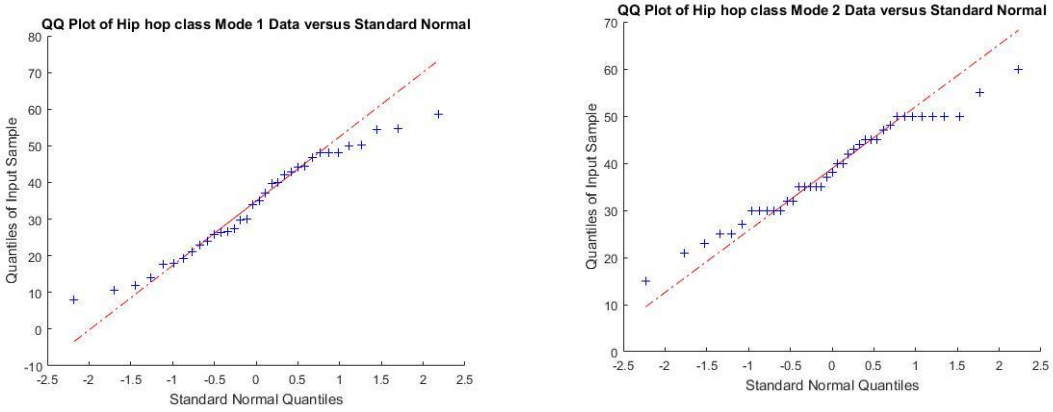
(a) Histograms. Elicited 50th percentile (median) from the hip hop class question.



(b) Boxplots. Elicited 50$^{th}$ percentile (median) from the hip hop class question.




(c) Quantile-quantile plots (Q-Q Plots) indicate the normality assumption holds.




**I.3 Comparison of the central values gathered from the basketball attendance question.**

We considered the elicited median values from the basketball attendance question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

In the case of the online elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the

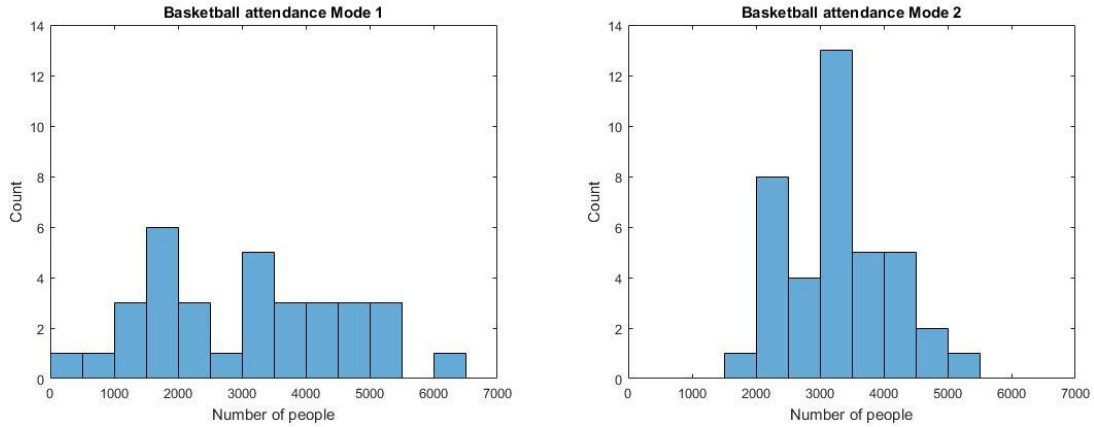
Shapiro-Wilk test for normality($SW = 0.97, p = .5681$), as well as the skewness $(0.11)$ and kurtosis $(-1.07)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.97, p = .274$), as well as the skewness $(0.32)$ and kurtosis $(-0.54)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution is normal.

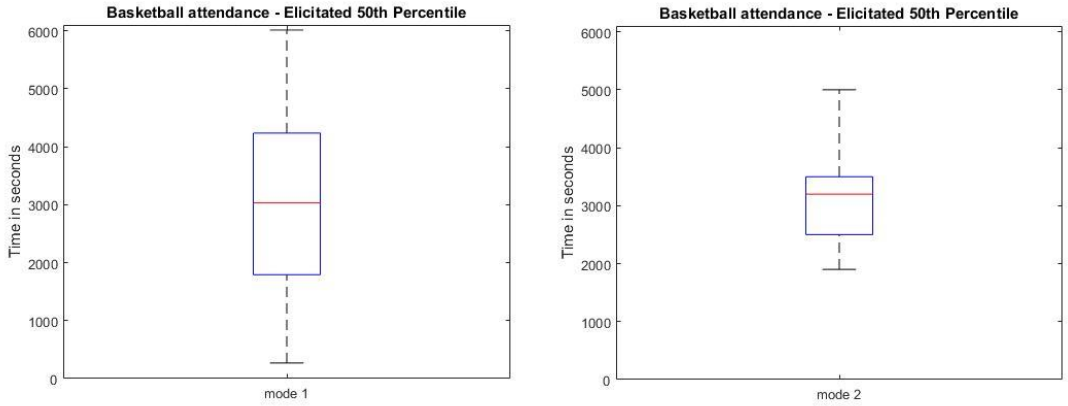
The underlying distributions were normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(45) = -0.44, p = .6647$ (two-sided); $g = -0.11$ (negligible). This result indicated no significant difference between the online mean elicited median values ($M = 3048, SD = 1493, n = 33$) and the face-to-face mean elicited median values ($M = 3173, SD = 739, n = 39$).

**Figure 18.** Comparison of the central values gathered from the basketball attendance question.

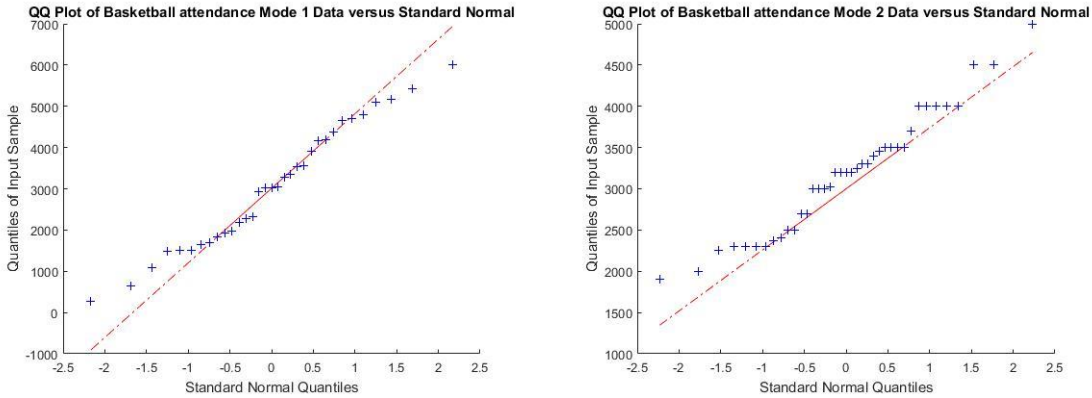
(a) Histograms. Elicited 50th percentile (median) from the basketball attendance question.




(b) Boxplots. Elicited 50th percentile (median) from the basketball attendance question.




(c) Quantile-quantile plots (Q-Q Plots) indicate the normality assumption holds.



**I.4 Comparison of the central values gathered from the Game of Thrones question.** We considered the elicited median values from the Game of Thrones question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 17 was removed from the analysis because their probability distribution was incomplete.

In the case of the online elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality($SW = 0.94\,, p = .08596$), as well as the skewness $(0.77)$and kurtosis $(0.17)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was negatively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.74\,, p < .001$), as well as the skewness $(-2.69)$ and kurtosis $(10.05)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

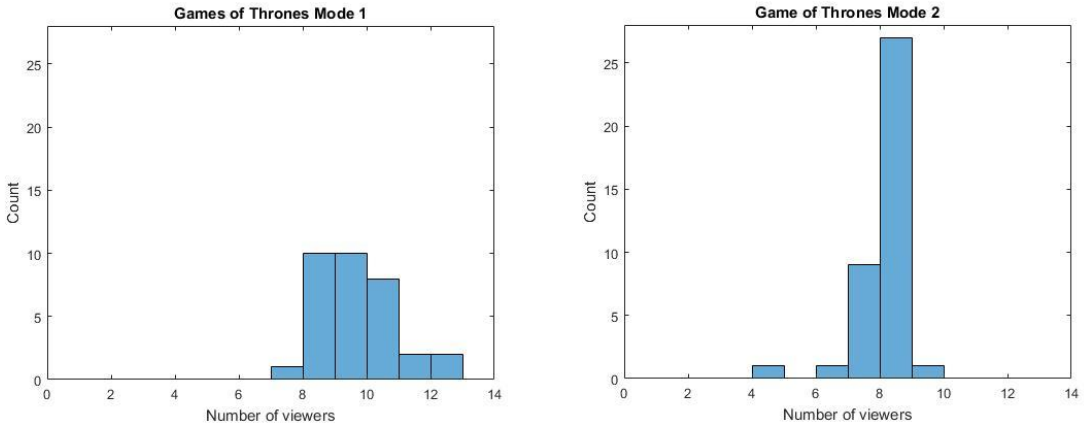
Although the underlying distributions from the face-to-face data was non-normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(50) = 4.98, p < .001$ (two-sided). This result indicated a significant difference between the online mean elicited median values ($M =$

9.22, $SD = 1.17$ , $n = 33$) and the face-to-face mean elicited median values ($M =$

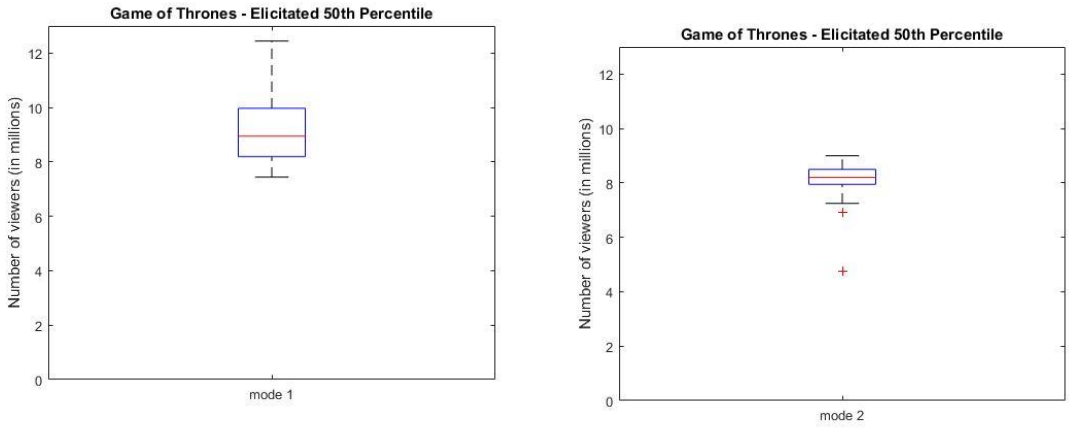
8.07, $SD = 0.7$ , $n = 39$).

**Figure 19.** Comparison of the central values gathered from the Game of Thrones question.

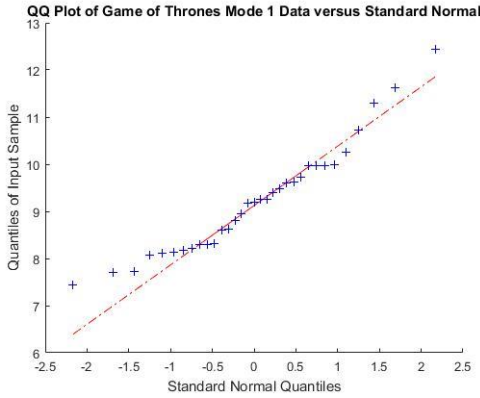
(a) Histograms. Elicited 50th percentile (median) from the Game of Thrones question.

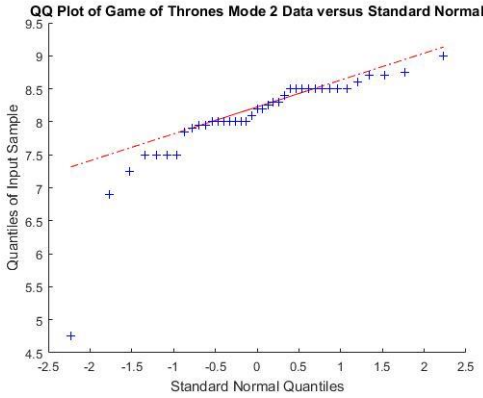



(b) Boxplots. Elicited 50th percentile (median) from the Game of Thrones question.



(c) Quantile-quantile plots (Q-Q Plots).




**I.5 Comparison of the central values gathered from the YouTube question.**

We considered the elicited median values from the YouTube question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

In the case of the online elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality($SW = 0.73 , p < .001$ ), as well as the skewness $(1.21)$and kurtosis $(0.01)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

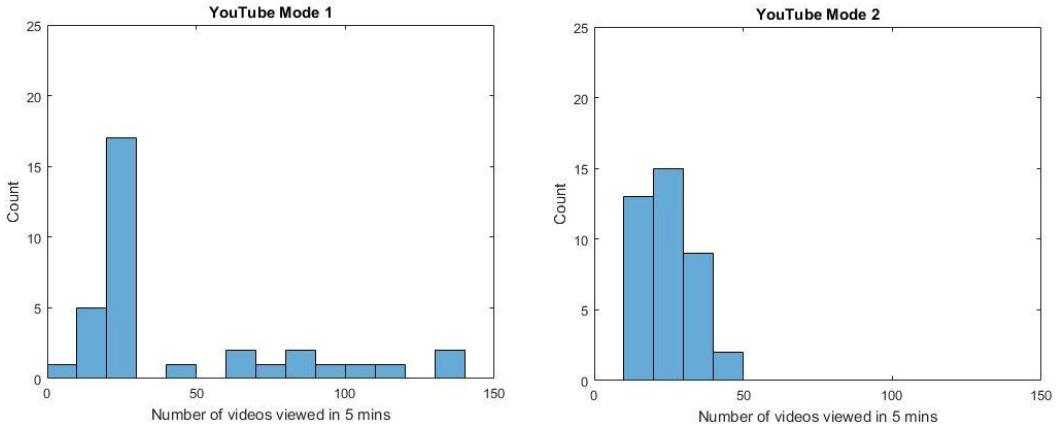
Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.83 , p < .001$), as well as the skewness $(1.38)$ and kurtosis $(1.48)$ statistics suggested the underlying distribution was not

normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

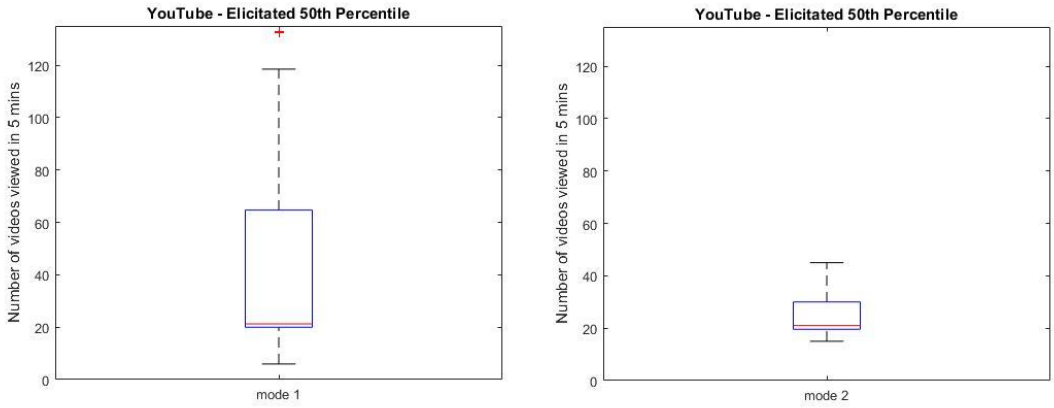
Although the underlying distributions were non-normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05, t(35) = 2.85, p = .007304$ (two-sided). This result indicated a significant difference between the online mean elicited median values ($M = 43, SD = 37, n = 34$) and the face-to-face mean elicited median values ($M = 247, SD = 7, n = 39$).

**Figure 20.** Comparison of the central values gathered from the YouTube question.

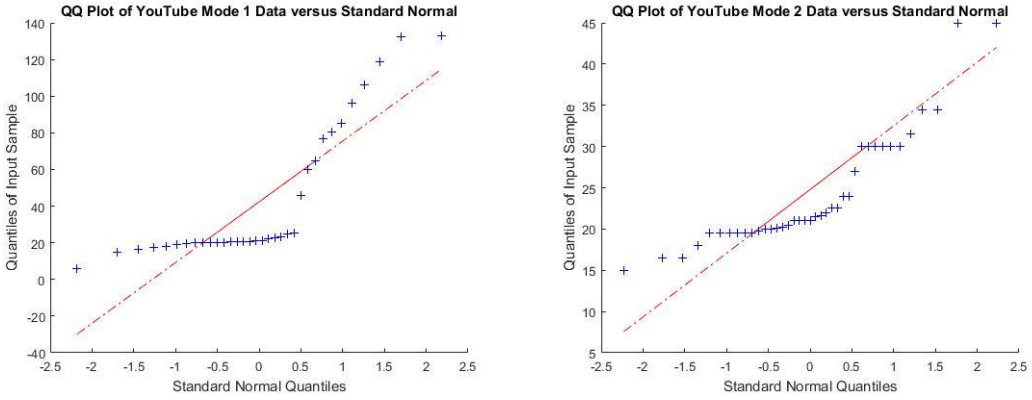
(a) Histograms. Elicited 50th percentile (median) from the YouTube question.



(b) Boxplots. Elicited 50th percentile (median) from the YouTube question.




(c) Quantile-quantile plots (Q-Q Plots).




**I.6 Comparison of the central values gathered from the opening weekend question.**

We considered the elicited median values from the opening weekend question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

In the case of the online elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality($SW = 0.946, p = .2769$), as well as the skewness

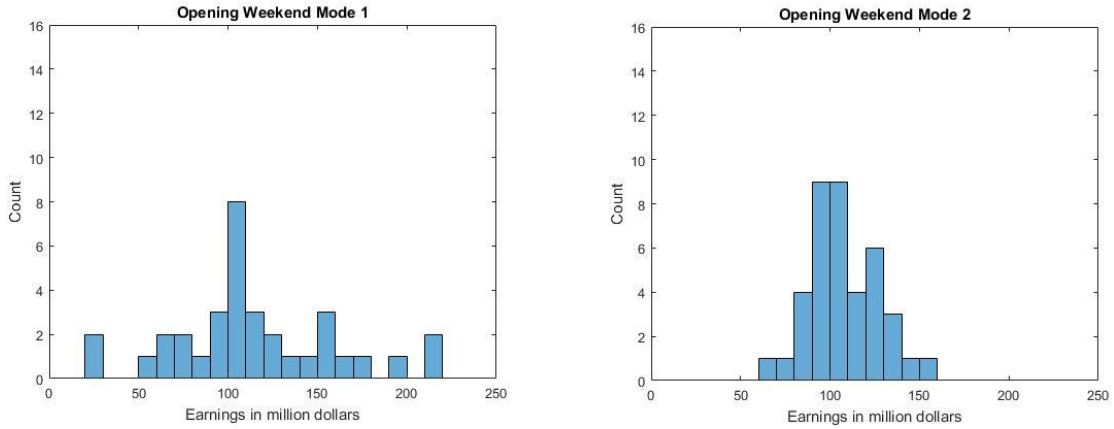
$(0.28)$and kurtosis $(-0.05)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.798, p = .7834$), as well as the skewness $(0.26)$ and kurtosis $(-0.02)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

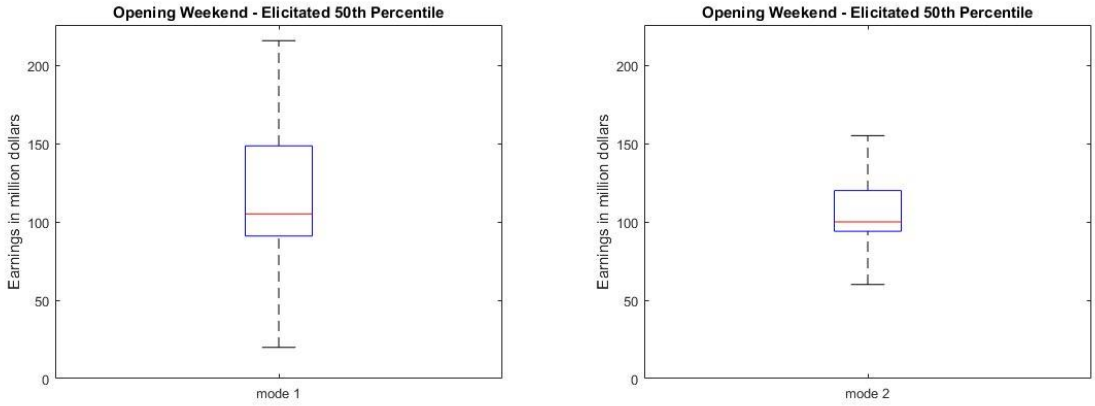
The underlying distributions were normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(42) = 0.92, p = 0.3637$ (two-sided); $g = 0.22$ (small). This result indicated no significant difference between the online mean elicited median values ($M = 113, SD = 46, n = 34$) and the face-to-face mean elicited median values ($M = 105, SD = 19, n = 39$).

**Figure 21.** Comparison of the central values gathered from the opening weekend question.

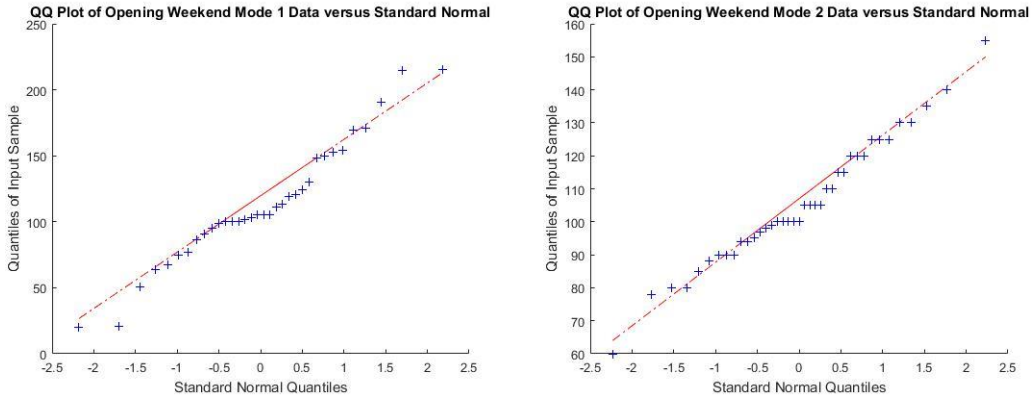
(a) Histograms. Elicited 50th percentile (median) from the opening weekend question.




(b) Boxplots. Elicited 50th percentile (median) from the opening weekend question.




(c) Quantile-quantile plots (Q-Q Plots).



**I.7 Comparison of the central values gathered from the high temperature question.** We considered the elicited median values from the high temperatures question. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 17, 33 and 69 were removed from the analysis because their probability distribution were incomplete.

In the case of the online elicited median values, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality($SW = 0.88\,, p = .002433$), as well as the skewness $(1.02)$and kurtosis $(0.12)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

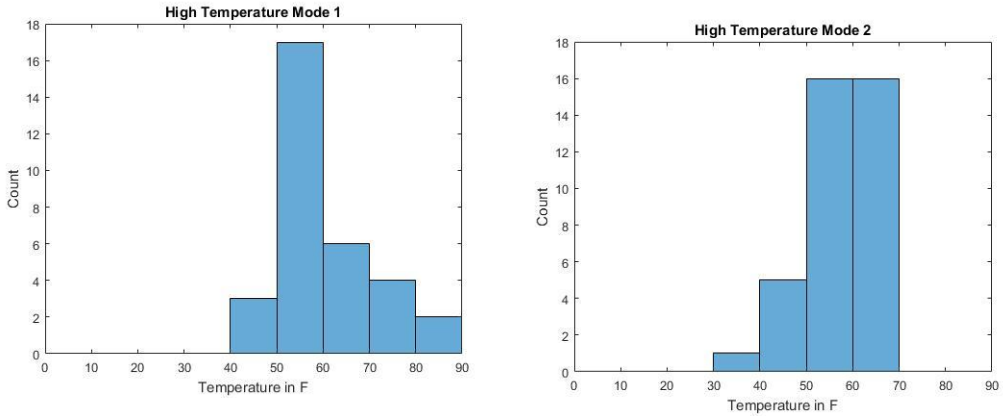
Regarding the face-to-face elicited median values, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.95\,, p = .09366$), as well as the skewness $(-0.49)$ and kurtosis $(0.41)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Although the underlying distributions from the online elicitation was non-normal, and next we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(56) = 1.83, p = .0733$ (two-sided). This result

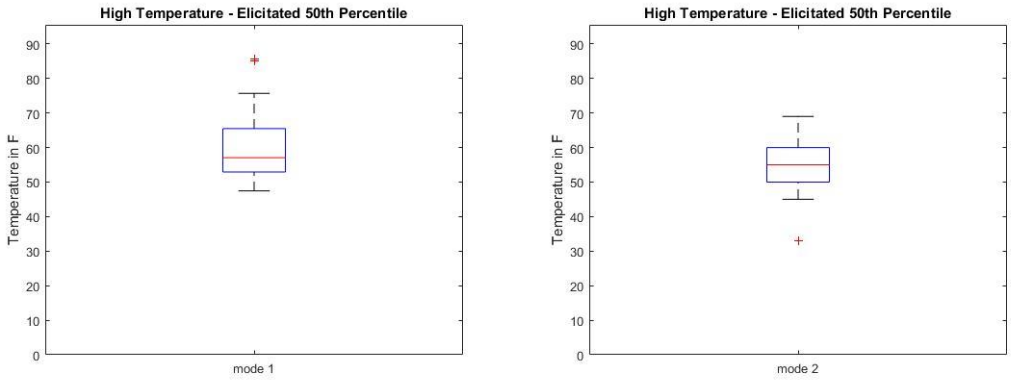
indicated no significant difference between the online mean elicited median values ($M = 60, SD = 10\ , n = 32$) and the face-to-face mean elicited median values ($M = 56, SD = 8\ , n = 39$).

**Figure 22.** Comparison of the central values gathered from the high temperature question.

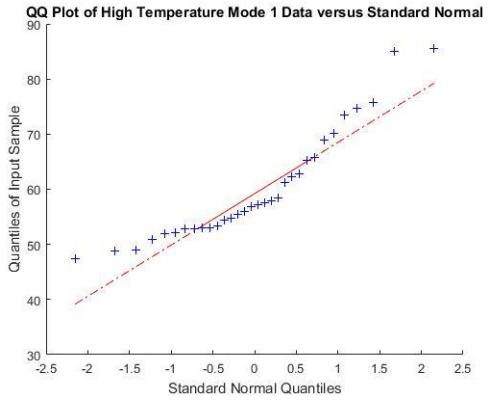
(a) Histograms. Elicited 50th percentile (median) from the high temperature question.

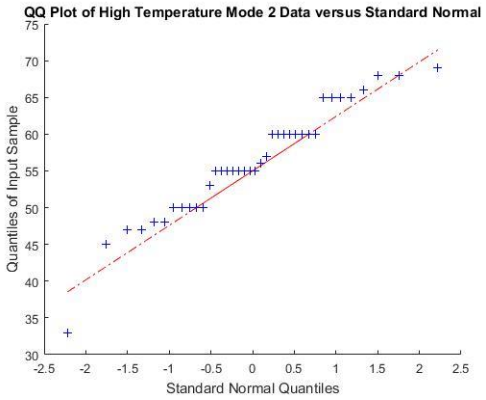



(b) Boxplot. Elicited 50th percentile (median) from the high temperature question.



(c) Quantile-quantile plots (Q-Q Plots).



## APPENDIX J

## COMPARISON OF THE UNCERTAINTY RANGE

We analyzed the uncertainty range over the sample space for question group 1. While not statistically significant, the mean online uncertainty range was higher than the face-to-face. Before we carried out an independent two sample $t$-test, we cleaned the data (Appendix F) and tested the assumption of normality.

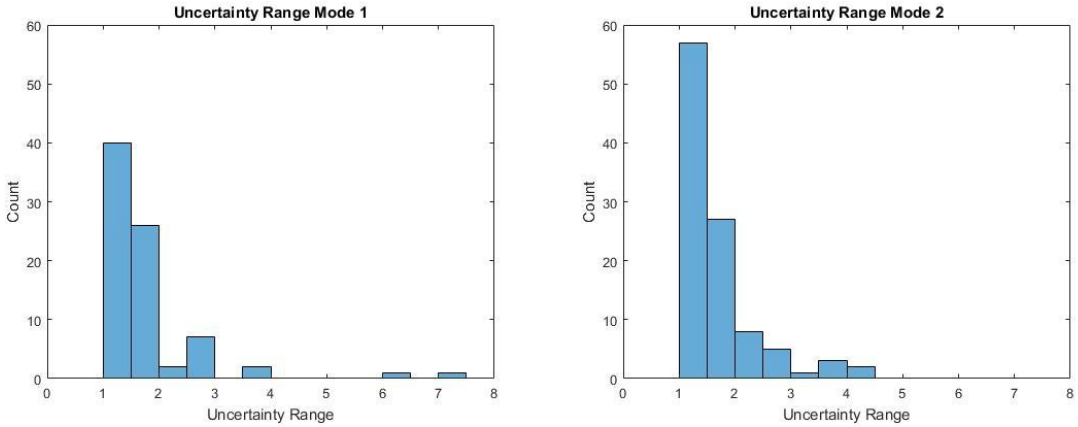
In the case of the online uncertainty range, the histogram suggested the underlying distribution was positively skewed. Review of the Shapiro-Wilk test for normality ($SW = 0.72, p < .001$), as well as the skewness $(3.38)$and kurtosis $(18.83)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Regarding the face-to-face uncertainty range, the histogram suggested the underlying distribution was positively skewed. Review of the Shapiro-Wilk test ($SW = 0.82, p < .001$), as well as the skewness $(2.07)$ and kurtosis $(6.11)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

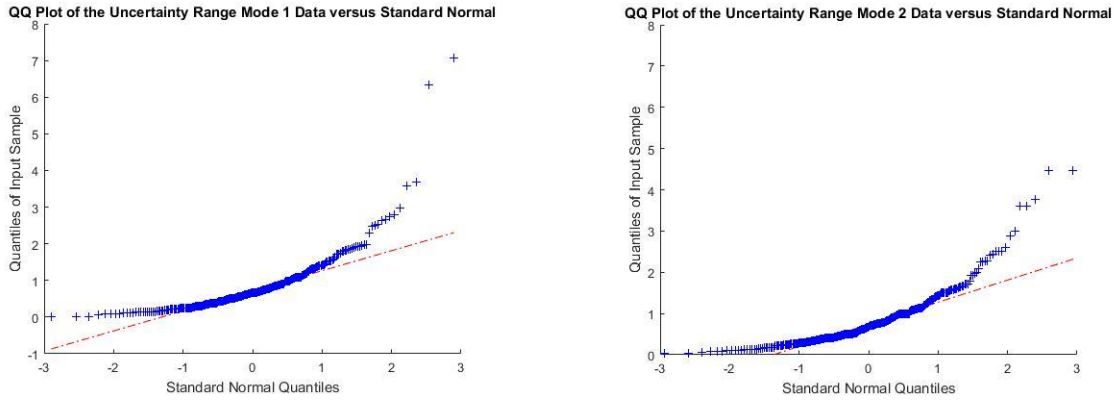
Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(524) = 0.17, p = .4326$ (one-sided). This result indicated the mean online uncertainty

range ($M = 0.86, SD = 0.83\,, n = 267$) was not significantly higher than the face-to-face mean uncertainty range ($M = 0.85, SD = 0.7\,, n = 309$).

**Figure 23.** Comparison of the uncertainty range.
(a) Histogram. Uncertainty range for question in group 1.




(b) Quantile-quantile plots (Q-Q Plots) indicate the normality assumption does not hold.



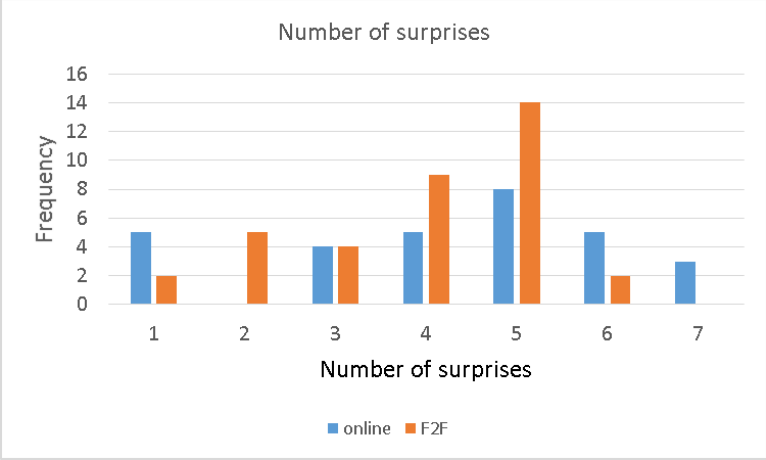
## APPENDIX K

## PROPORTION OF SURPRISES

**Table 13.** Summary table of the proportion of surprises.

| | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | Total |
|---|---|---|---|---|---|---|---|---|
| Online | $\frac{12}{30}$ | $\frac{18}{30}$ | $\frac{12}{30}$ | $\frac{19}{30}$ | $\frac{23}{30}$ | $\frac{23}{30}$ | $\frac{21}{30}$ | $\frac{128}{210}$ |
| F2F | $\frac{3}{36}$ | $\frac{16}{36}$ | $\frac{16}{36}$ | $\frac{32}{36}$ | $\frac{24}{36}$ | $\frac{20}{36}$ | $\frac{31}{36}$ | $\frac{142}{252}$ |
| Total | $\frac{15}{66}$ | $\frac{34}{66}$ | $\frac{28}{66}$ | $\frac{51}{66}$ | $\frac{47}{66}$ | $\frac{43}{66}$ | $\frac{52}{66}$ | |

**Figure 24.** Number of surprises. The frequency table takes account of the sample space for question group 1. On average, the online and face-to-face are surprised 5 out of 7 times (when considering the modal number of surprises). Note the sample sizes are unequal.



## APPENDIX L

## CORE SCORES FROM QUESTIONS IN GROUP 1

**Table 14.** Individual experts' core scores for each question in group 1.
*Incomplete subjective probability distributions (with either one or more elicited value missing), therefore no core score assigned.
**Elicited median estimate equal to zero, therefore no core score assigned.

**Online Core Scores**

| **Expert** $(i)$ | $k=1$ $A_{i\,1\,1}$ | $k=2$ $A_{i\,1\,2}$ | $k=3$ $A_{i\,1\,3}$ | $k=4$ $A_{i\,1\,4}$ | $k=5$ $A_{i\,1\,5}$ | $k=6$ $A_{i\,1\,6}$ | $k=7$ $A_{i\,1\,7}$ |
|---|---|---|---|---|---|---|---|
| 1 | 2224.89 | 3188.862 | 59903.42 | 58.2399 | 127.3046 | 4850.525 | 1362.865 |
| 2 | 2792.439 | 5337.514 | 45304.9 | 53.38998 | 286.7443 | 7474.467 | 3884.763 |
| 3 | 7213.458 | 944.971 | 118424.3 | 136.1061 | 3504.699 | 2812.663 | 1233.203 |
| 4 | 1036.901 | 714.9003 | 284628.2 | 48.56489 | 5996.695 | 9834.152 | 753.4083 |
| 5 | 2209.033 | 528.904 | 93211.72 | 73.83001 | 179.7293 | 443.3795 | 1269.069 |
| 6 | 2949.016 | 3354.851 | 39019.48 | 225.033 | 186.3454 | 4334.591 | 890.0757 |
| 7 | 881.1205 | 2237.687 | 93339.81 | 87.74421 | 28.8939 | 4056.028 | 1661.539 |
| 8 | 1979.439 | 6793.657 | 139706.2 | 36.00838 | 9028.773 | 2104.543 | 3326.832 |
| 9 | 6000.367 | 4603.333 | 60700.91 | 225.2366 | 120.584 | 1544.081 | 579.0757 |
| 10 | 7420.6 | 4970.457 | 52152.21 | 14.29155 | 10407.73 | 4248.519 | 2647.594 |
| 11 | 7929.767 | 984.4899 | 23527.94 | 27.0864 | 129.2928 | 11789.85 | 2874.059 |
| 12 | 426.7431 | 3915.112 | 55721.72 | 178.1898 | 137.4425 | 4029.046 | 2275.522 |
| 13 | 1358.484 | 1083.005 | 55431.99 | 104.7988 | 168.1982 | 2909.46 | 1363.266 |
| 14 | 1967.91 | 4932.09 | 40048.46 | 90.82653 | 12966 | 264.7762 | 448.9307 |
| 15 | 1152.217 | 3457.644 | 71600.68 | 20.39689 | 56.27209 | 1122.949 | 279.7019 |
| 16 | 10005.12 | 727.4978 | * | 154.2328 | 8065.469 | 15204.61 | 2323.212 |
| 17 | 4178.581 | 6102.049 | 20318.62 | * | 382.7993 | 5156.44 | * |
| 18 | 2598.067 | 3534.431 | 362190.7 | 54.59101 | 97.12736 | 7286.981 | 3246.603 |
| 19 | 1248.92 | 1596.92 | 53280.34 | 36.40496 | 321.5426 | 1591.261 | 1824.716 |
| 20 | 1221.285 | 1103.183 | 74663.33 | 122.4649 | 211.3935 | 1818.26 | 1085.799 |
| 21 | 3572.486 | 35.65939 | 160428.7 | 17.0152 | 1817.757 | 7014.367 | 2478.247 |
| 22 | * | 5474.649 | 259473.7 | 57.48413 | 47.03998 | 4827.358 | 2393.927 |
| 23 | 1372.305 | 1785.634 | 26318.4 | 218.953 | 200.0658 | 2880.858 | 2430.218 |
| 24 | 6868.289 | 72.63867 | 301415.8 | 224.915 | 13752.65 | 667.3158 | 3393.863 |
| 25 | 2614.023 | 2499.627 | 126602.5 | 195.7033 | 77.64641 | 2164.829 | 1966.173 |
| 26 | 1224.944 | 217.9696 | 105490.6 | 44.93312 | 316.066 | 4609.068 | 1446.046 |
| 27 | 795.4444 | 221.135 | 109838.8 | 391.6837 | 448.5197 | 13810.18 | 633.0716 |
| 28 | 2371.163 | 2811.94 | 50739.57 | 221.039 | 128.7414 | 3950.074 | 1427.779 |
| 29 | 3691.098 | 467.7315 | 314921 | 27.20792 | 2516.379 | 3337.787 | 1556.817 |
| 30 | 3005.995 | 524.7025 | 33886.88 | 116.6725 | 79.08325 | 1497.567 | 977.1191 |
| 31 | 2877.06 | 2775.496 | 344477.9 | 194.6243 | 70.35545 | 7727.837 | 2221.343 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 32 | 15179.2 | 2586.456 | 119788.1 | 320.9788 | 14707.25 | 6066.494 | 1021.709 |
| 33 | 23020.18 | 1945.907 | 94400.4 | 53.68453 | 3491.371 | 2689.314 | * |
| 34 | 1146.402 | 1346.157 | 158875 | 154.9685 | 1520.924 | 708.0264 | 3424.426 |

**F2F Core Scores**

| Expert ($i$) | $k=1$ $A_{i21}$ | $k=2$ $A_{i22}$ | $k=3$ $A_{i23}$ | $k=4$ $A_{i24}$ | $k=5$ $A_{i25}$ | $k=6$ $A_{i26}$ | $k=7$ $A_{i27}$ |
|---|---|---|---|---|---|---|---|
| 35 | 2070 | 3915 | 73300 | 166.5 | 341.565 | 5075.5 | 2025 |
| 36 | 1725 | 2425 | 68400 | 104 | 91.5645 | 1400.5 | 1160 |
| 37 | 3400 | 2575 | 102400 | 196.5 | 78.435 | 7176.5 | 1550 |
| 38 | 2200 | 3300 | 121400 | 247.5 | 1675.305 | 1931.5 | 1560 |
| 39 | 850 | 2300 | 18300 | 162.5 | 603.435 | 1526.5 | 1975 |
| 40 | 3450 | 2300 | 77400 | 301.5 | 1443.435 | 6126.5 | 740 |
| 41 | 465 | 1875 | 41700 | 124 | 693.435 | 2075.5 | 2300 |
| 42 | 8550 | 325 | 54900 | 216.5 | 161.565 | 3926.5 | 2700 |
| 43 | 815 | 1315 | 143300 | 119.5 | 618.435 | 3075.5 | 1515 |
| 44 | 2000 | 1760 | 52650 | 323 | 65.565 | 7676.5 | 2395 |
| 45 | 1800 | 2415 | 102400 | 186.5 | 580.935 | 4651.5 | 2575 |
| 46 | 3140 | 3245 | 92300 | 196.8 | 153.435 | 1676.5 | 2630 |
| 47 | 3450 | 2650 | 48300 | 144.5 | 71.574 | 6626.5 | 2825 |
| 48 | 8670 | 3015 | 23110 | 211.5 | 52.065 | 7676.5 | 2840 |
| 49 | 2025 | 1225 | 15700 | 154 | 63.4245 | 5401.5 | 1975 |
| 50 | 1875 | 3275 | 53400 | 231.5 | 356.565 | 2276.5 | 2440 |
| 51 | 1815 | 1025 | 40800 | 113 | 693.435 | 4125.5 | 2575 |
| 52 | 1375 | 785 | 20555 | 303.5 | 26.565 | 6876.5 | 3740 |
| 53 | 2475 | 1935 | 132400 | 184 | 81.5655 | 8751.5 | 950 |
| 54 | 7250 | 5905 | 20800 | 81.85 | 36.663 | 2675.5 | 2360 |
| 55 | 5115 | 1150 | 144900 | 70.5 | 476.565 | 3726.5 | 3450 |
| 56 | 2700 | 4660 | 3450 | 134 | 140.973 | 7276.5 | 4480 |
| 57 | ** | 3085 | 37300 | 203 | 271.5645 | 2751.5 | 910 |
| 58 | 1215 | 4800 | 15800 | 91.5 | 341.565 | 1075.5 | 1775 |
| 59 | ** | 3100 | 72300 | 141.5 | 1570.935 | 5176.5 | 3085 |
| 60 | 1385 | 2725 | 20300 | 458.1 | 814.185 | 3196.5 | 1405 |
| 61 | 1850 | 2725 | 27200 | 249.8 | 54.075 | 6876.5 | 1300 |
| 62 | 3450 | 2625 | 73300 | 102.5 | 1008.435 | 2976.5 | 1475 |
| 63 | 875 | 3500 | 49800 | 296.5 | 139.065 | 7976.5 | 2905 |
| 64 | 4200 | 4075 | 93300 | 327.5 | 243.429 | 1925.5 | 3725 |
| 65 | 1095 | 1900 | 62400 | 271.5 | 228.435 | 1605.5 | 2950 |
| 66 | 1150 | 3125 | 209900 | 110.5 | 310.935 | 6251.5 | 1825 |
| 67 | 3450 | 4250 | 10700 | 149 | 70.935 | 4851.5 | 1475 |
| 68 | 3500 | 5000 | 42400 | 181.5 | 96.564 | 774.5 | 2600 |
| 69 | 7200 | 3160 | 104900 | 272 | 811.4355 | 6236.5 | * |
| 70 | 3720 | 2425 | 164900 | 321 | 116.565 | 7676.5 | 1285 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 71 | 1000 | 675 | 31700 | 174 | 335.934 | 855.5 | 865 |
| 72 | 615 | 2100 | 30700 | 111.5 | 625.935 | 2675.5 | 950 |
| 73 | 2000 | 2425 | 109900 | 310.5 | 48.435 | 475.5 | 3590 |

# APPENDIX M

## COMPARISON OF THE ACCURACY OF THE FORECASTS

The analysis of the difference in mean core scores was approached question by question.

### M.1 Comparison of the accuracy of the forecasts from the library elevator question

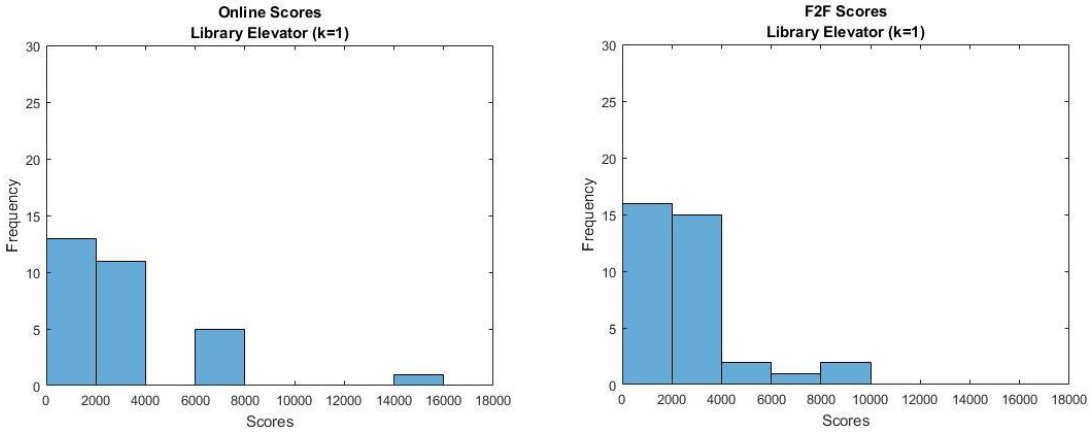
We considered the core scores from the library elevator question. While not statistically significant, the face-to-face mean core score was lower than the online. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 22, 57 and 59 were removed from the analysis because their probability distributions were incomplete.

In the case of the online core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality($SW = 0.68, p < .001$), as well as the skewness $(2.46)$ and kurtosis $(6.48)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

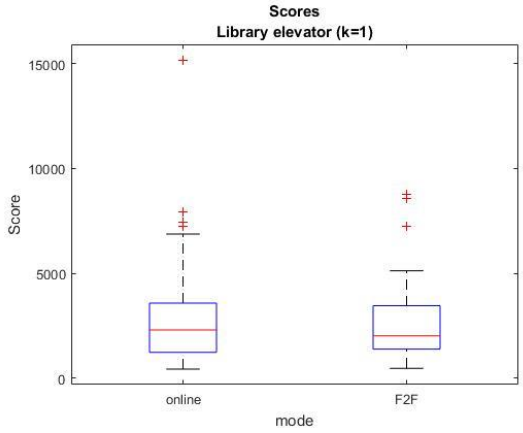
Regarding the face-to-face core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.82, p < .001$), as well as the skewness $(1.44)$ and kurtosis $(1.34)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution is non-normal.

Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(43) = 1.42, p = .08144$ (one-sided). This result indicated the online mean core score ($M = 4077,\ SD = 4654\ , n = 33$) was not significantly higher than the face-to-face mean core score ($M = 2824, SD = 2120\ , n = 37$).

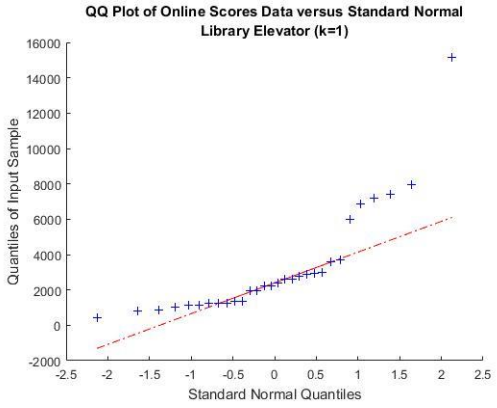
**Figure 25.** Comparison of the core scores gathered from the library elevator question. (a) Histogram. Core scores from the library elevator question.

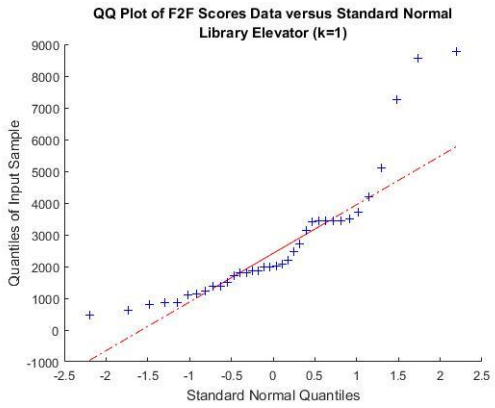



(b) Boxplot. Core scores from the library elevator question.



(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does not hold.




**M.2 Comparison of the accuracy of the forecasts from the hip hop class question.**

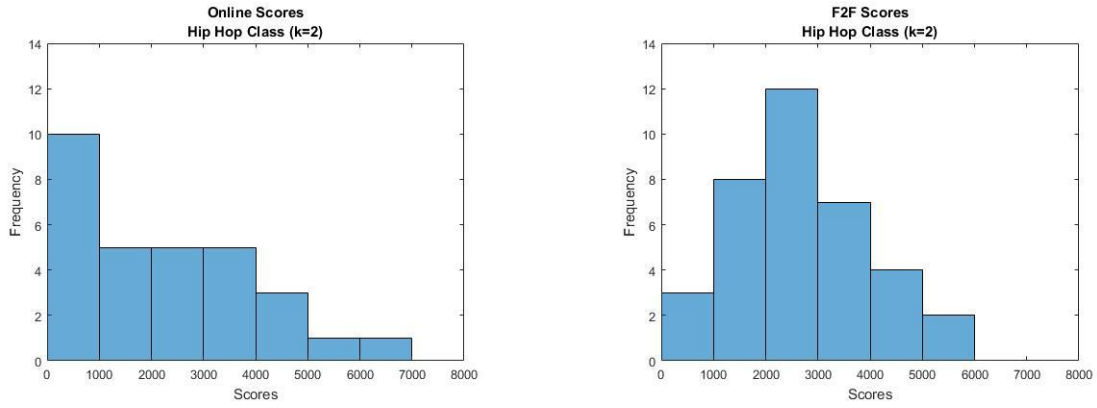
We consider the core scores from the hip hop class question. While not statistically significant, the online mean core score was lower than the face-to-face. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

In the case of the online core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.93\,, p = .02305$), as well as the skewness $(0.57)$ and kurtosis $(-0.87)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

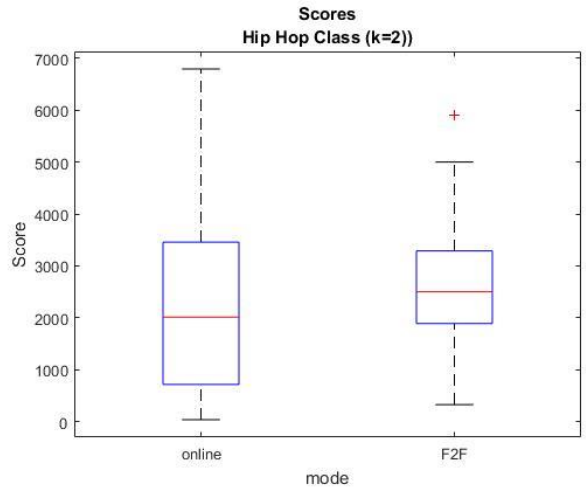
Regarding the face-to-face core scores, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.98, p = .6577$), as well as the skewness $(0.36)$ and kurtosis $(-0.11)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Although the underlying distribution for the online core scores was non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(54) = 0.67, p = .2542$ (one-sided). This result indicated the face-to-face mean core score ($M = 2694, SD = 1241, n = 39$) was not significantly higher than online mean core score ($M = 2438, SD = 1925, n = 34$).

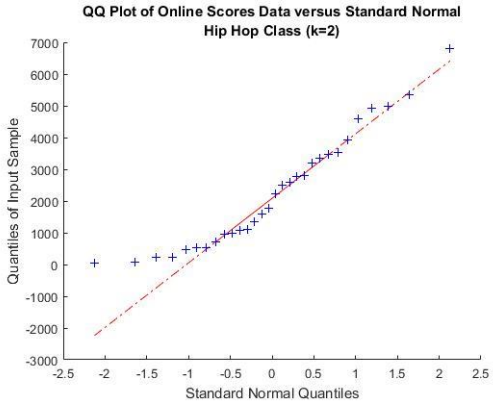
**Figure 26.** Comparison of the core scores gathered from the hip hop class question. (a) Histograms. Core scores from hip hop class question.

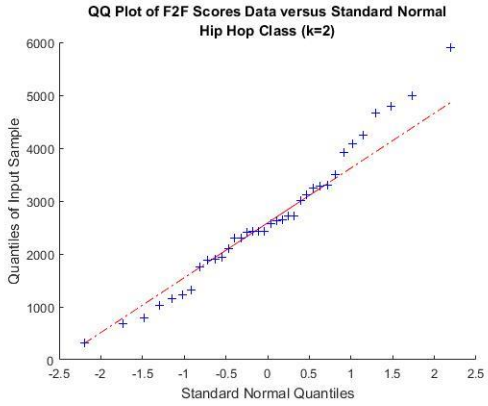


(b) Boxplots. Core scores from hip hop class question.




(c) Quantile-quantile plots (Q-Q plots).




**M.3 Comparison of the accuracy of the forecasts from the basketball attendance question.**

We considered the core scores from the basketball attendance question. The face-to-face mean core score was lower than the online suggesting the face-to-face forecasts were more accurate. Before we carried out an independent two sample $t$- test, we cleaned the

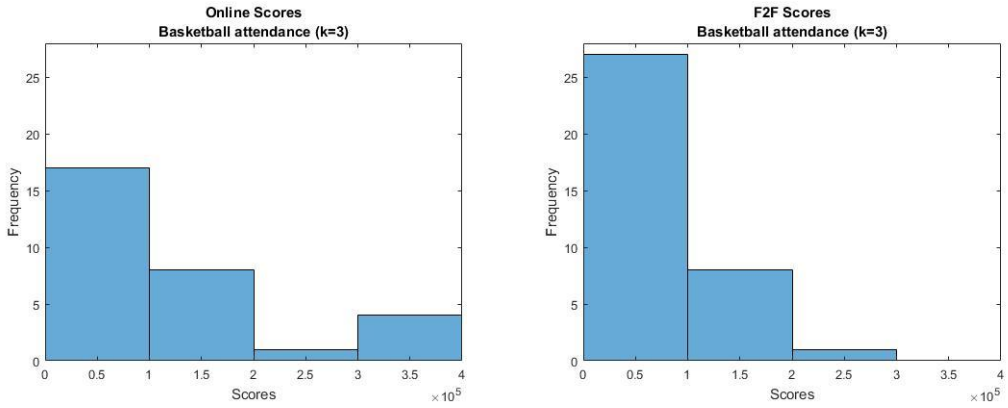
data and tested the assumption of normality. Expert 16 was removed from the analysis because their probability distribution was incomplete.

In the case of the online core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.81, p < .001$), as well as the skewness $(1.2)$and kurtosis $(0.1)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

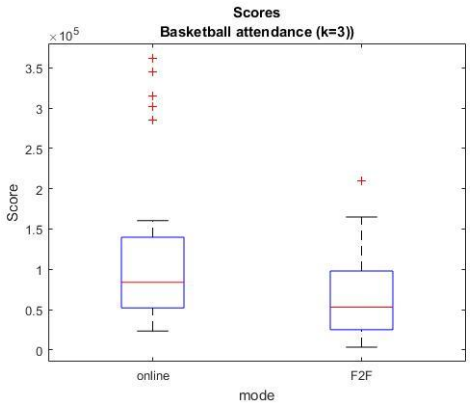
Regarding the face-to-face core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.92\,, p = .008123$ ), as well as the skewness $(0.93)$ and kurtosis $(0.31)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically significant at $\alpha = .05$, $t(44) = 2.77, p = .0040541$ (one sided). This result indicated the online mean core score ($M = 119692,\ SD = 100020, n = 33$) was significantly higher than the face-to-face mean core score ($M = 66889,\ SD = 48084\,, n = 39$).

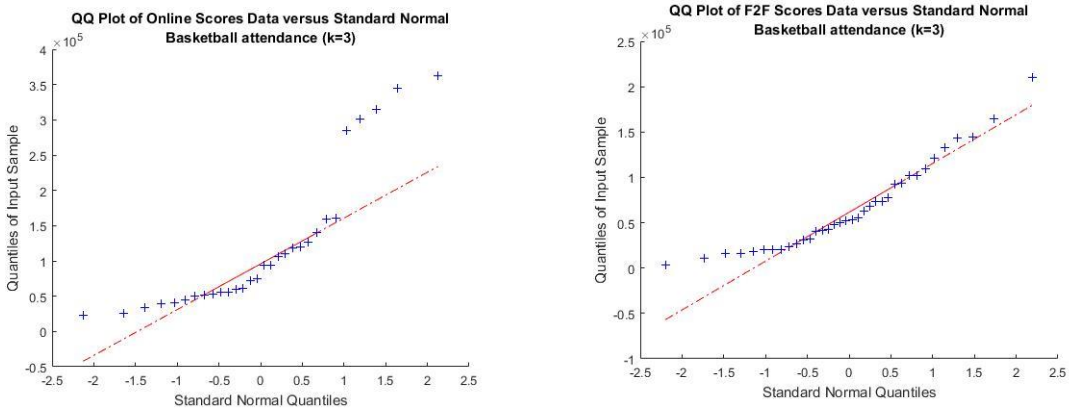
**Figure 27.** Comparison of the core scores gathered from the basketball attendance question.
(a) Histograms. Core scores from basketball attendance question.




(b) Boxplots. Core scores from basketball attendance question.




(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does not hold.



**M.4 Comparison of the accuracy of the forecasts from the Game of Thrones question.**

We considered the core scores from the Game of Thrones question. The online mean core score was lower than the face-to-face suggesting the online forecasts were more accurate. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 17 was removed from the analysis because their probability distribution was incomplete.

In the case of the online core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.89, p = .003175$), as well as the skewness $(0.94)$ and kurtosis $(0.23)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

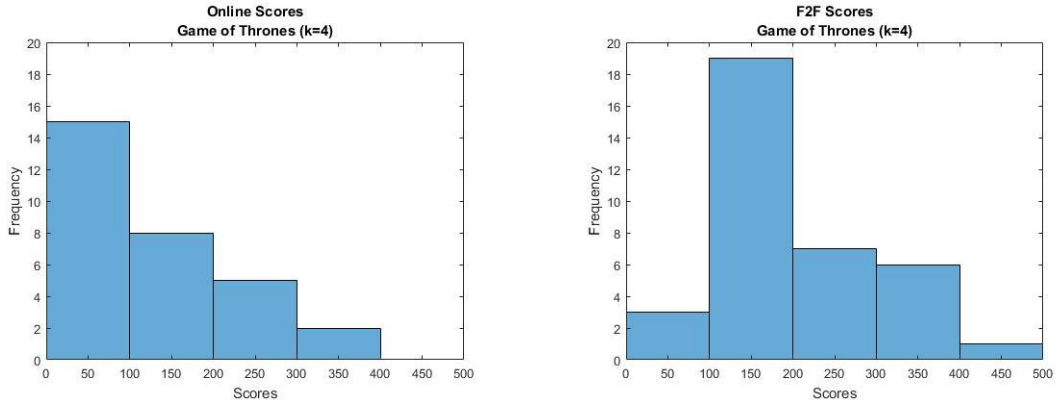
Regarding the face-to-face core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.94, p = .04125$), as well as the skewness $(0.72)$ and kurtosis $(0.07)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Although the underlying distributions are non-normal, we carry out an independent two sample $t$-test. This test was found to be statistically significant at $\alpha = .05$, $t(65) = 3.61, p < .001$ (one sided). This result indicated the face-to-face mean core score ($M =$

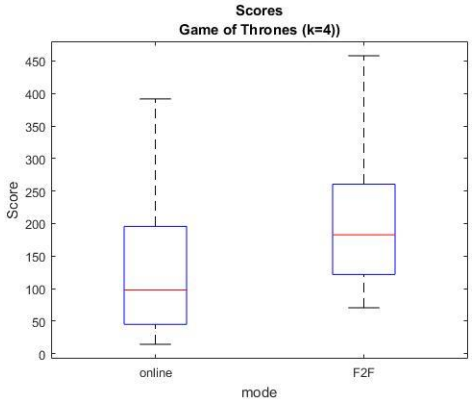
$199, SD = 87, n = 39$) was significantly higher than the online mean core score ($M = 121, SD = 94, n = 33$).

**Figure 28.** Comparison of the core scores gathered from the Game of Thrones question.

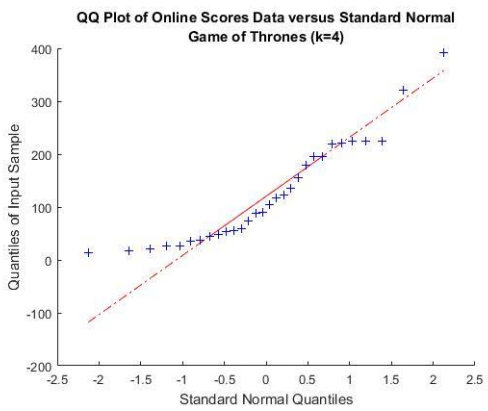
(a) Histograms. Core score from Game of Thrones question.

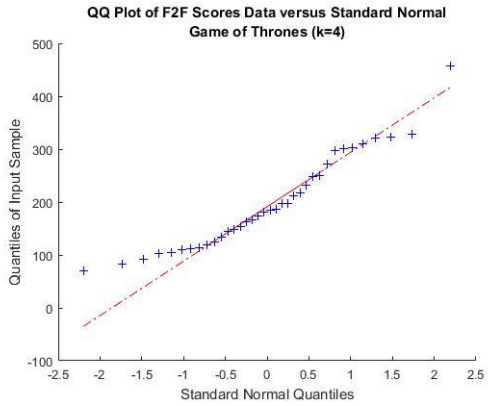



(b) Boxplots. Core score from Game of Thrones question.



(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does not hold.




**M.5 Comparison of the accuracy of the forecasts from the YouTube question.**

We considered the core scores from the YouTube question. The face-to-face mean core score was lower than the online suggesting the face-to-face forecasts were more accurate. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

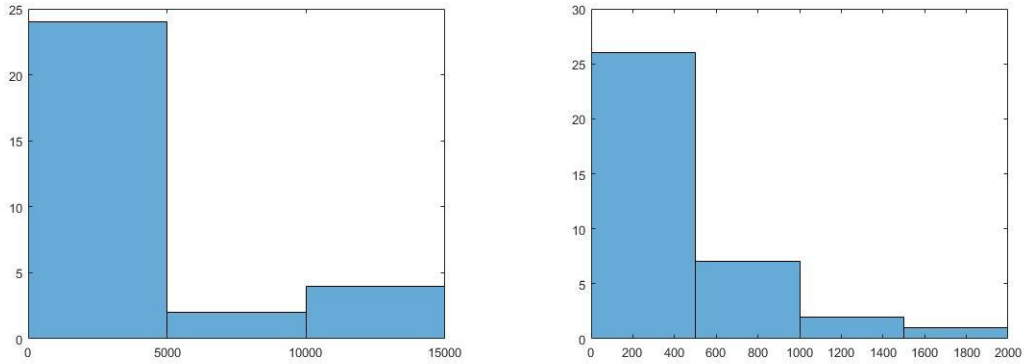
In the case of the online core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.65\,, p < .001$), as well as the skewness $(1.57)$and kurtosis $(1.03)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Regarding the face-to-face core scores, the histogram suggested the underlying distribution was positively skewed and the boxplot indicated outliers. Review of the Shapiro-Wilk test ($SW = 0.79\,, p < .001$), as well as the skewness $(1.5)$ and kurtosis

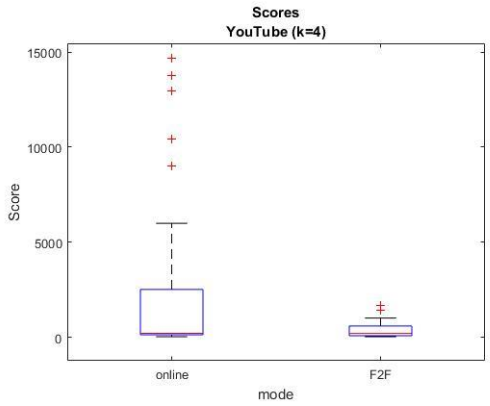
(1.57) statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically significant at $\alpha = .05$, $t(33) = 2.99, p = .002601$ (one sided). This result indicated the online mean core score ($M = 2693,\ SD = 4455, n = 34$) is significantly higher than the face-to-face core score ($M = 400,\ SD = 430\ , n = 39$).

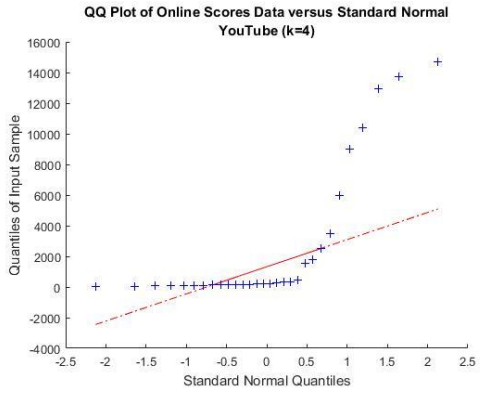
**Figure 29.** Comparison of the core scores gathered from the YouTube question.
(a) Histograms. Core scores from YouTube question.

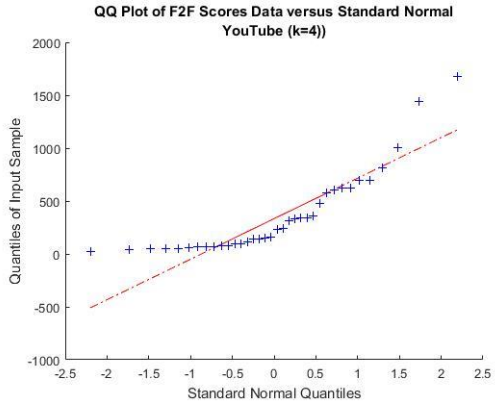


(b) Boxplots. Core scores from YouTube question.




(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does not hold.




**M.6 Comparison of the accuracy of the forecasts from the opening weekend question.**

We considered the core scores from the opening weekend question. While not statistically significant, the face-to-face mean core score was lower than the online. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. No experts were removed from the analysis.

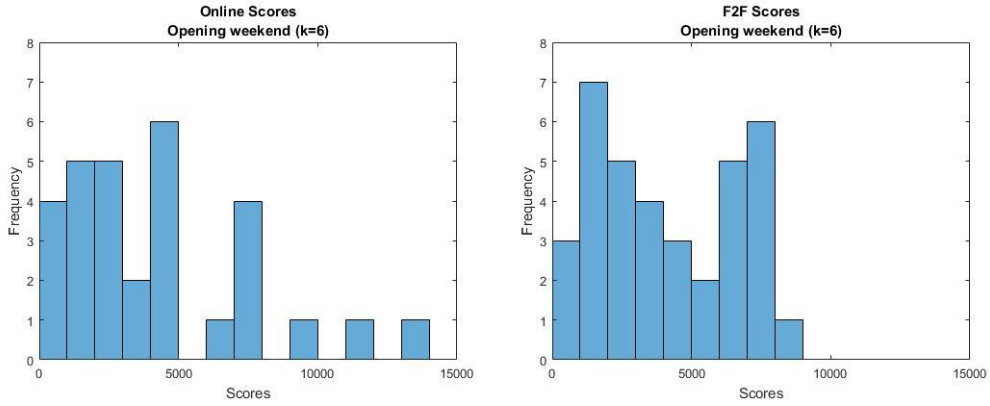
In the case of the online core scores, the histogram suggested the underlying distribution has two peaks and the boxplot indicated an outlier. Review of the Shapiro-Wilk test for normality($SW = 0.87, p < .001$), as well as the skewness $(1.26)$and kurtosis $(1.01)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

Regarding the face-to-face core scores, the histogram suggested the underlying distribution has two peaks and the boxplot indicates no outlier. Review of the Shapiro-Wilk test ($SW = 0.84, p = .02741$), as well as the skewness $(0.17)$ and kurtosis $(-1.41)$ statistics suggested the underlying distribution was not normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was non-normal.

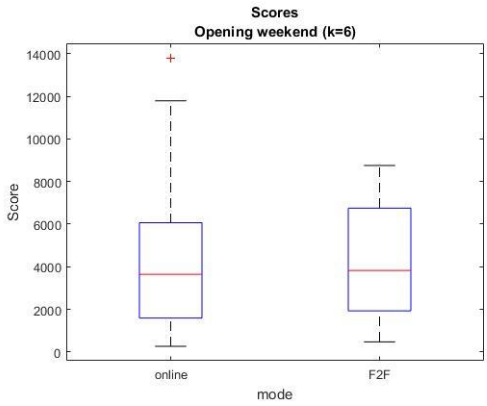
Although the underlying distributions were non-normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(56) = 0.43, p = .3347$ (one-sided). This result indicated the online mean core score ($M = 4554,\ SD = 3702, n = 34$) was not significantly higher than the face-to-face mean core score ($M = 4233,\ SD = 2458, n = 39$).

**Figure 30.** Comparison of the core scores gathered from the opening weekend question.

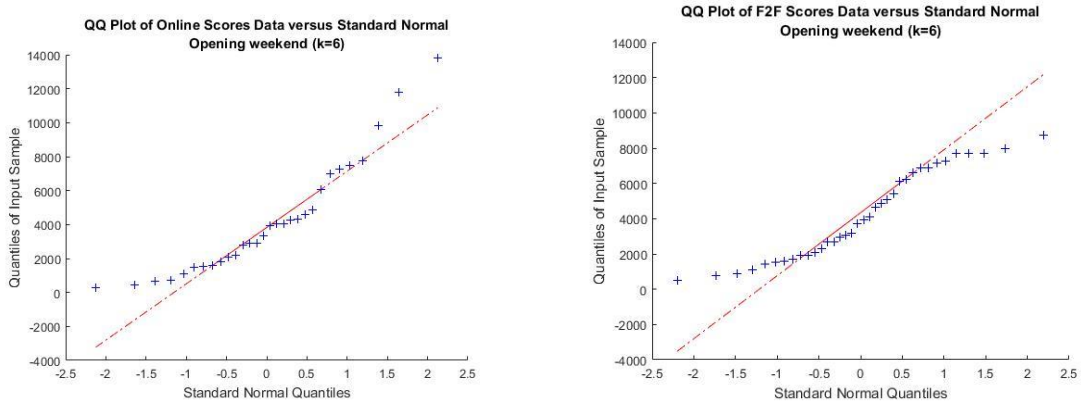
(a) Histograms. Core scores from the opening weekend question.




(b) Boxplots. Core scores from the opening weekend question.




(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does hold.



**M.7 Comparison of the accuracy of the forecasts from the high temperature question.**

We considered the core scores from the high temperature question. While not statistically significant, the online mean core score was lower than the face-to-face. Before we carried out an independent two sample $t$-test, we cleaned the data and tested the assumption of normality. Expert 17, 33 and 69 were removed from the analysis because their probability distributions were incomplete.

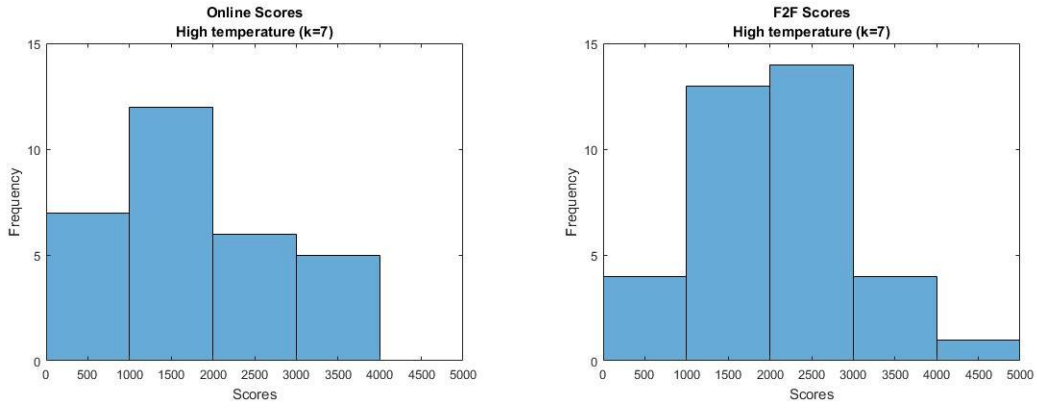
In the case of the online core scores, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test for normality ($SW = 0.96, p = .2529$), as well as the skewness $(0.35)$and kurtosis $(-1)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

Regarding the face-to-face core scores, the histogram suggested the underlying distribution was symmetrical and the boxplot indicated no outliers. Review of the Shapiro-Wilk test ($SW = 0.96, p = .2707$), as well as the skewness $(0.38)$ and kurtosis $(-0.66)$ statistics suggested the underlying distribution was normal. The Q-Q plot confirmed the result and we concluded the underlying distribution was normal.

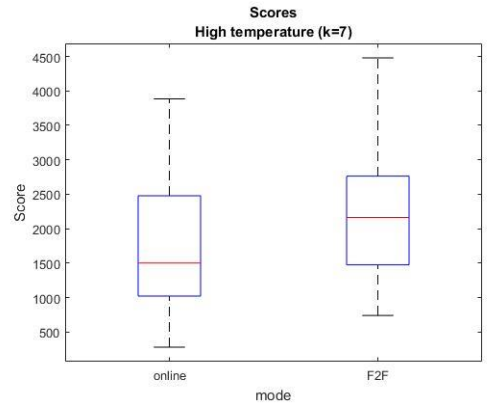
Since the underlying distributions were normal, we carried out an independent two sample $t$-test. This test was found to be statistically non-significant at $\alpha = .05$, $t(64) = 1.51, p = .06788$ (one sided). This result indicated the face-to-face mean core score

($M = 2181, SD = 928, n = 38$) was not significantly higher than the online mean core score ($M = 1834, SD = 979\,, n = 32$).

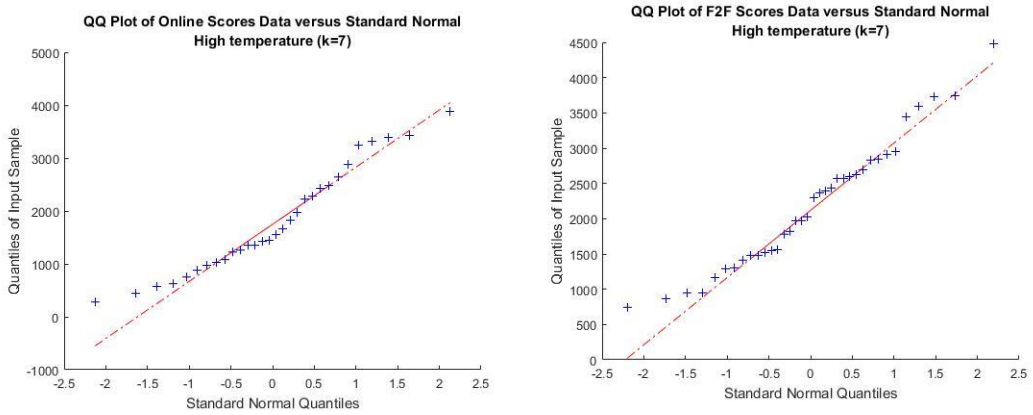
**Figure 31.** Comparison of the core scores gathered from the high temperature question. (a) Histogram. Core scores from the high temperature question.




(b) Boxplot. Core scores from the high temperature question.



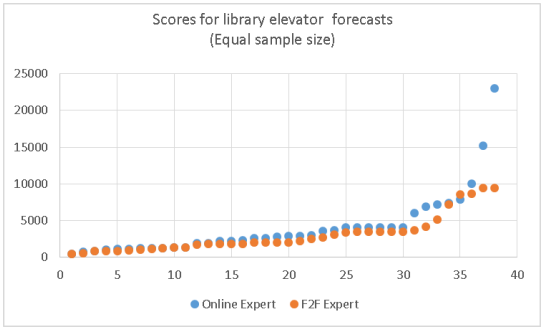

(c) Quantile-quantile plots (Q-Q plots) indicate the normality assumption does hold.

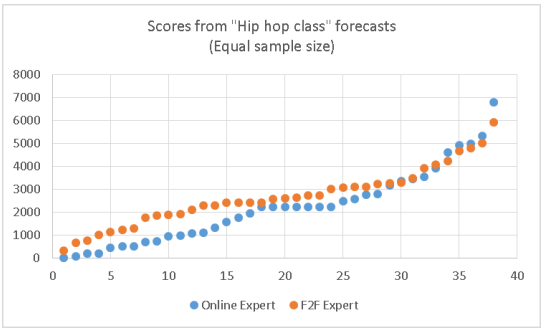


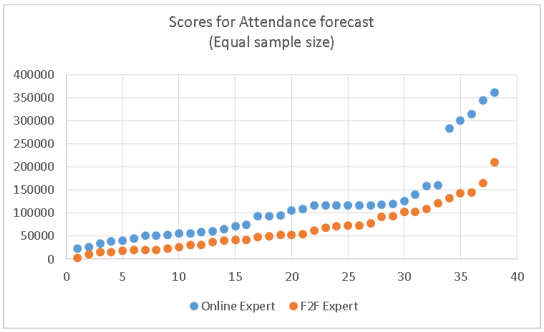
## APPENDIX N

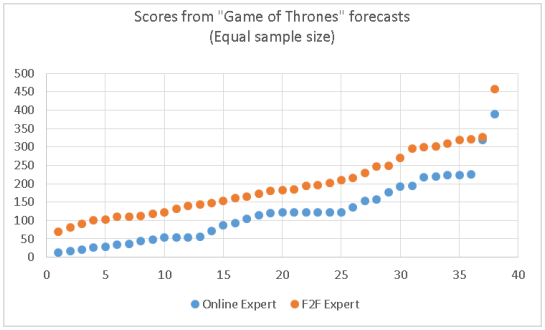
## SCATTER PLOTS OF CORE SCORES

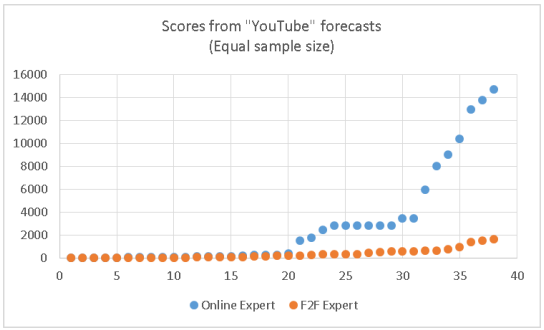
**Figure 32.** Scatter plots of core scores. The data is adjusted to force equal sample size. Forty core scores are displayed for each survey mode. In the instances where the sample size is less than 40, additional scores equal to the mean are included.

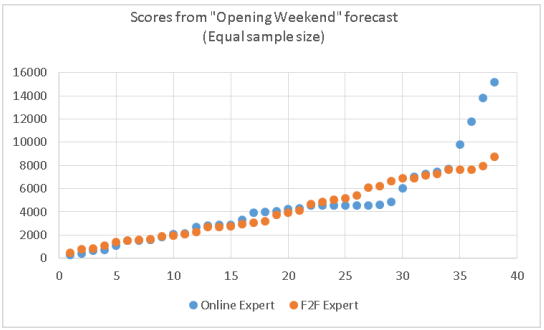


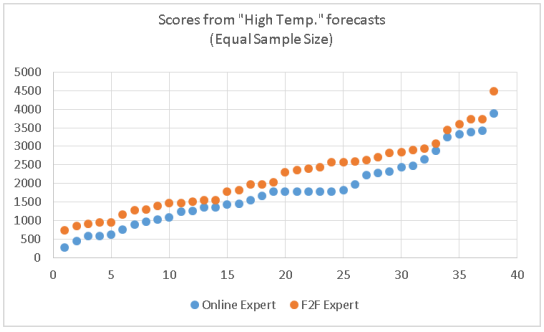



## APPENDIX O

## TIME TAKEN TO COMPLETE THE ONLINE ELICITATIONS

The average time taken to complete the online elicitation was 1 hour 45 minutes (105 minutes). However, 24 online elicitation were completed within 1 hour. Expert 27 took the shortest time to complete the online elicitation (11 minutes). The online elicitation also gathered written feedback. Out of the 19 opportunities to provide written feedback, expert 27 gave 17 responses; two of which were "not sure".

**Table 15.** Summary table of the time taken to complete the online elicitation.

| **Time taken to complete online elicitation** | |
|---|---|
| Mean | 1 hour 45 minutes (105 minutes) |
| Trimmed mean (10%) | 1 hour 12 minutes (72 minutes) |
| Min | 11 minutes |
| Max | 20 hours 48 minutes (1249 minutes) |

**Table 16.** Frequency table of the time taken to complete the online elicitation.

| **Time interval (minutes)** | **Number of elicitations completed within time interval** |
|---|---|
| 0 - 30 | 12 |
| 31 - 60 | 12 |
| 61 - 90 | 6 |
| 91 - 120 | 1 |
| More than 120 | 3 |

**Table 17.** Summary table of the time taken to complete each online elicitation question.

| Time taken to respond to a particular question (minutes) | | | | | |
|---|---|---|---|---|---|
| **Question** | **Mean** | **Median** | **Mix** | **Max** | **Trimmed Mean (10%)** |
| 1 | 1.8 | 1.4 | 0.2 | 7.5 | 1.7 |
| 2 | 2.2 | 1.4 | 0.3 | 8.9 | 2.1 |
| 3 | 11.4 | 1.3 | 0.4 | 332.9 | 1.7 |
| 4 | 1 | 0.8 | 0.2 | 5.1 | 0.9 |
| 5 | 2.8 | 1.7 | 0.3 | 30.8 | 2 |
| 6 | 2 | 1.3 | 0.2 | 19.7 | 1.5 |
| 7 | 3 | 1.7 | 0.3 | 21.1 | 2.6 |
| 8 | 8 | 1.1 | 0.2 | 209.7 | 2 |
| 9 | 4.1 | 1.6 | 0.3 | 39.2 | 3.1 |
| 10 | 4 | 1.2 | 0.3 | 85.6 | 1.5 |
| 11 | 28.2 | 1.3 | 0.3 | 905.9 | 1.7 |
| 12 | 2.4 | 1.5 | 0.4 | 12 | 2.2 |
| 13 | 2.5 | 1.8 | 0.3 | 11.3 | 2.3 |
| 14 | 3.2 | 1.3 | 0.3 | 47.2 | 1.9 |
| 15 | 16.8 | 1.5 | 0.3 | 499 | 2.2 |
| 16 | 2.9 | 1.4 | 0.4 | 33.8 | 2.1 |
| 17 | 1.8 | 1.2 | 0.2 | 6.3 | 1.7 |
| 18 | 2.1 | 1.3 | 0.3 | 17.4 | 1.7 |
| 19 | 1.4 | 1.2 | 0.1 | 6 | 1.3 |
| 20 | 2.9 | 2 | 0.2 | 15.4 | 2.6 |

## APPENDIX P

## TIME TAKEN TO COMPLETE THE FACE-TO-FACE ELICITATIONS

At the beginning of the face-to-face elicitation, the interviewer gave an overview of the study as detailed in the consent form. Following on, participants had an opportunity to ask questions about the study before signing the consent form. The times measured below do not include this introduction section.

**Table 18.** Summary table of the time taken to complete the face-to-face elicitation.

| **Time taken to complete online elicitation** | |
|---|---|
| Mean | 1 hour 31 minutes (91 minutes) |
| Trimmed mean (10%) | 1 hour 31 minutes (91 minutes) |
| Median | 1 hour 30 minutes (90 minutes) |
| Min | 1 hour 4 minutes (64 minutes ) |
| Max | 2 hours 3 minutes (123 minutes) |

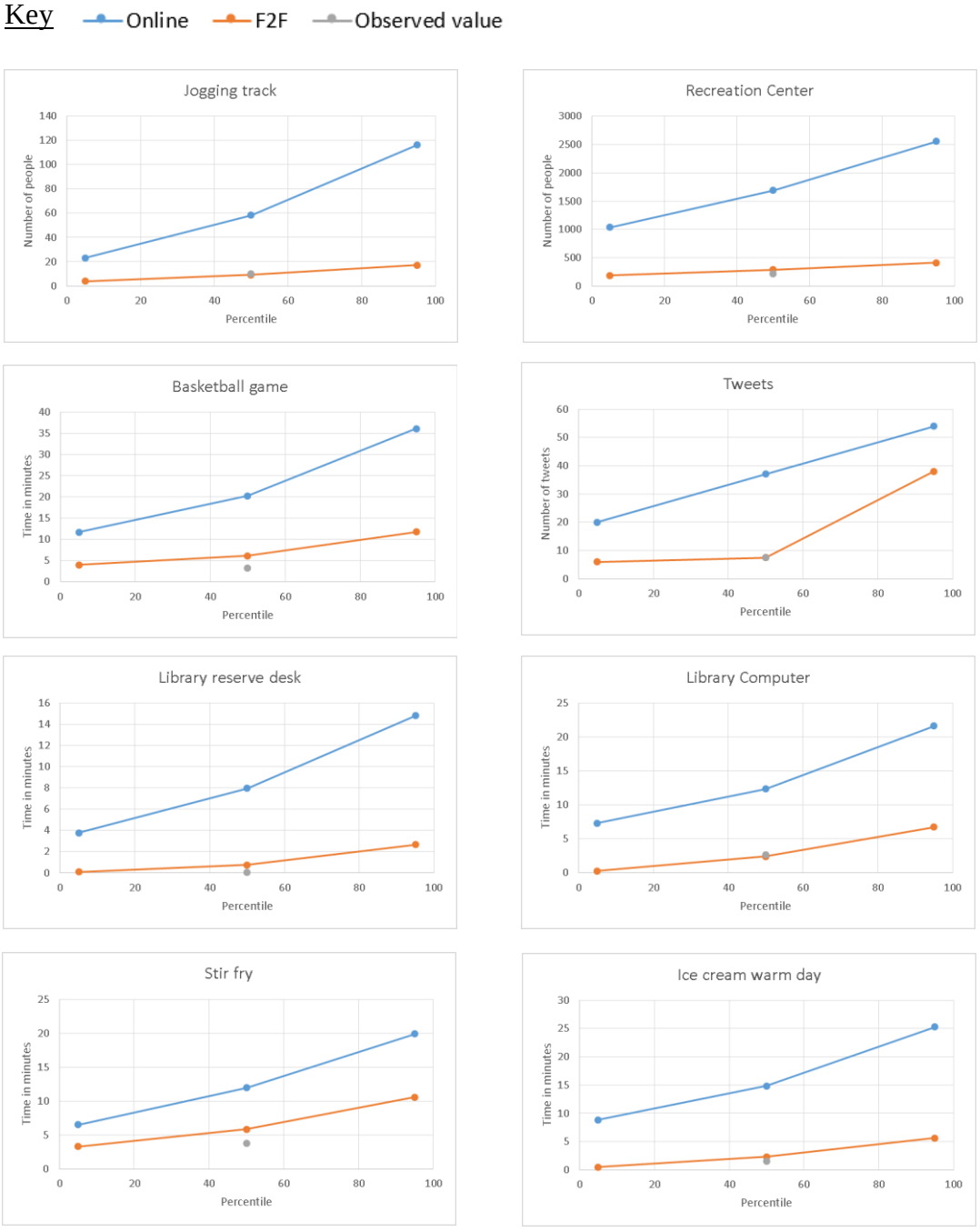
**Table 19.** Frequency table of the time taken to complete the face-to-face elicitation.

| **Time interval (minutes)** | **Number of elicitations completed within time interval** |
|---|---|
| 0 - 30 | 0 |
| 31 - 60 | 0 |
| 61 - 90 | 18 |
| 91 - 120 | 15 |
| More than 120 | 2 |

## APPENDIX Q

## AGGREGATED PROBABILITY DISTRIBUTIONS FOR QUESTIONS IN GROUP 2 AND 3

**Figure 33.** The aggregated probability distributions for each question in group 2 and 3.



## APPENDIX R

## COMPARISON OF ACCURACY FOR QUESTIONS IN GROUP 2 AND 3

**Table 20.** Which mode resulted in more accurate values? Results from the independent two sample $t$-tests.

| Question | More accurate? | Significantly lower score at 5% level? | Online | | F2F | | $t$ | $df$ | $p$ (one-sided) | Normality assumption holds? |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | M | SD | M | SD | | | | |
| Jogging track | F2F | No (Yes at 6%) | 4531 | 15319 | 229 | 163 | 1.62 | 32 | 0.05827 | No |
| Rec center | F2F | Yes | 165589 | 244544 | 7645 | 6429 | 3.71 | 32 | 0.000393 | No |
| Basketball game | F2F | Yes | 1859 | 2093 | 285 | 196 | 4.37 | 33 | <0.0001 | No |
| Tweets | F2F | Yes | 3101 | 4265 | 192 | 306 | 3.91 | 32 | 0.000224 | No |
| Library reserve desk | F2F | Yes | 824 | 1068 | 56 | 63 | 4.12 | 32 | 0.000124 | No |
| Library computer | F2F | Yes | 997 | 1367 | 173 | 112 | 3.45 | 32 | 0.000789 | No |
| Stir fry | F2F | Yes | 824 | 891 | 169 | 185 | 4.21 | 35 | <0.0001 | No |
| Ice cream warm day | F2F | Yes | 1096 | 936 | 425 | 228 | 4.07 | 36 | 0.000121 | No |

## APPENDIX S

## ANALYSIS OF OPEN ENDED FACE-TO-FACE QUESTION

We analyzed 39 experts' open responses from the face-to-face elicitation to the Game of Thrones question. Participants were not limited to give one discrete reason and some participants thought of multiple. The Game of Thornes question asked how many people would tune in for the Season 7 premiere of HBO's *Game of Thrones* telecast in spring of 2017. Participants were asked: "Suppose the number of viewers turned out to be higher than your high estimate, why would that happen?"

Responses to this question were grouped into eight categories: additional advertising; a lot of new viewers; interesting plot; HBO reduced subscription or special offer; bad weather; recommended by a friend; not competing with other shows.

## BIBLIOGRAPHY


Anadon, L. D., Nemet, G. & Verdolini, E. (2013). The future costs of nuclear power using multiple expert elicitation: effects of RD&D and elicitation design. Environmental Research Letters, 8(3).

Aspinall, W. P., Cooke, R. M., Havelaar, A. H., Hoffmann, S. & Hald, T. (2016). Evaluation of a Performance-Based Expert Elicitation: WHO Global Attribution of Foodborne Diseases. *PLoS ONE*, 11(3). doi: 10.1371/journal.pone.0149817

Baker, E (2016). Does the elicitation mode matter? Comparing different methods for eliciting expert judgment. Alfred P. Sloan Foundation grant proposal, The University of Massachusetts, Amherst.

Bartlett, M. S. (1947). The Use of Transformations. *Biometrics*, 3(2), 39-52.

Barton, B. & Peat, J. (2014). Medical statistics: a guide to SPSS, data analysis, and critical appraisal. Oxford, England: Wiley Blackwell.

Bistline, J. (2013). Energy technology expert elicitations: An application to natural gas turbine efficiencies. *Technological Forecasting & Social Change*, 86, 177-187.

Bowling, A. (2005). Mode of Questionnaire Administration Can Have Serious Effects on Data Quality. Journal of Public Health, 27, 281–291.

Budescu, D. V., N. Du. (2007). Coherence and consistency of investors' probability judgments. *Management Science*, 53(11), 1731-1744.

Chan, G., Anadon, L. D., Chan, M. & Lee, A. (2011). Expert Elicitation of Cost, Performance, and RD&D Budgets for Coal Power with CCS. *Energy Procedia*, 4, 2685-2692.

Clemen, R. T. and Reilly, T. (2001). Making hard decisions with Decision Tools. Pacific Grove: Duxbury Thomson Learning.

Clemen, R. T. & Winkler, R. L. (1999). Combining Probability Distribution From Experts in Risk Analysis. *Risk Analysis*, 19(2), 187-203.

Curtright, A, Morgan, M. G., Keith, D. (2008). Expert assessment of future photovoltaic technology. *Environmental Science and Technology*, 42, 9031-9038.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

Gaba, A., Tsetlin, I. & Winkler, R. L. (2017). Combining Interval Forecasts. *Decision Analysis*, 14(1), 1-20.

Garthwaite, P. H., Kadane, J. B., O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, 100(470), 680-701.

Ghasemi, A. & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology & Metabolism*, 10(2), 486-489.

Grushka-Cockayne, Y., Lichtendahl, K. C., Jose, V. R. R. & Winkler, R. L. (2017). Quantile Evaluation, Sensitivity to Bracketing, and Sharing Business Payoffs. *Operations Research*, 65(3), 712-728.

Hyndman, R. and Koehler A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679-688.

Inman, M. & Davis, S. J. (2012). How Low Will Solar Photovoltaic Prices Go? Summary of a Near Zero Expert Elicitation. (March 11th, 2018). Retrieved from: http://www.nearzero.org/reports/pv-learning/

Inman, M. & Davis, S. J (2012). Energy High in the Sky. Expert Perspectives on Airborne Wind Energy Systems. (March 11th, 2018) Retrieved from: http://www.nearzero.org/reports/AirborneWind

Ioannou, I., Aspinall, W., Rush, D., Bisby, L. & Rossetto, T. (2017). Expert judgment-based fragility assessment of reinforced concrete buildings exposed to fire. *Reliability Engineering and System Safety*, 167, 105-127.

Jain, A., Nandakumar, K. & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, 38, 2270-2285.

Joanes, D. N. & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician*, 47(1), 183-189

Jose, V. R. R., Winkler, R. L. (2008). Simple robust average of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1), 163–169.

Jose, V. R. R. & Winkler, R. L. (2009). Evaluating Quantile Assessments. *Operations Research*, 57(5), 1287-1297.

Komsta, L. & Novomestky, F. (2015, February 20). Package 'moments'. Cran R-projects. Retrieved from https://cran.r-project.org/web/packages/moments/moments.pdf

Krosnic, J. A. (1991). Response Strategies for Coping with Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213-236.

Lichtendahl, K. C., Grushka-Cockayne, Y. & Winkler, R. L. (2013). Is It Better to Average Probabilities or Quantiles?. Management Science, 59(7), 1594-1611.

Mann, P. (2010). Introductory Statistics (Seventh Edition). Danvers, MA: Wiley.

Marquard, J. L. & Robinson, S. M. (2008). Reducing perceptual and cognitive challenges in making decisions with models. In: Kugler, T., Smith, J. C., Connolly, T. & Son, Y. J. (Eds.) Decision Modeling and Behavior in Complex and Uncertain Environments. Springer Optimization and Its Applications, volume 21. Springer, New York, NY.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National Academy of Science of the Unites States of America, 111(20), 7176-7184.

Morgan, M. G. & Henrion, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, UK: Cambridge University Press.

Moser, B. K., Stevens, G. R. & Watts, C. L. (1989). The two-sample t test versus Satterthwaite's approximate f test. *Communications in Statistics - Theory and Methods*, 18(11), 3963-3975.

Nemet, G. F., Anadon, L. D. & Verdolini, E. (2017). Quantifying the Effects of Expert Selection and Elicitation Design on Experts' Confidence in Their Judgments About Future Energy Technologies. Risk Analysis, 37(2), 315-330.

Nordhaus, W. D. (1994). Expert opinion on climatic change. *American Scientist*, 82, 45–51.

Osborne, J. W. (2002). Notes on the use of data transformation. *Practical assessment, Research & Evaluation*, 8(6).

Peterson, C. R. & Miller, A. (1964). Mode, median and mean as optimal strategies. *Journal of Experimental Psychology*, 68(4), 363 -367.

Peterson, C. R., Snapper, K. J. & Murphy, A. H. (1972). Credible interval temperature forecasts. *Bulletin American Meteorological Society*, 53(10), 966-970.

Prava, V. R., Clemen, R. T., Hobbs, B. F., Kenney, M. A. (2016). Partition Dependence and Carryover Biases in Subjective Probability Assessment Surveys for Continuous Variables: Model-Based Estimation and Correction. *Decision Analysis*, 13 (1), 51-67.

Revelle, W. (2017, September 9). Package 'psych'. Cran R-projects. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Sharpe, P. & Keelin, T. (1998). How SmithKline Beecham makes better resource allocation decisions. Harvard Business Review, 76(2), 45–57.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52 (3/4), 591 -611.

Simon, H. A. (1972). Theories of bounded rationality. In McGuire, C. B. & Radner, R. (Eds.), Decision and Organization. A volume in honor of Jacob Marschak (p. 161-176). Amsterdam: North-Holland Publishing Company.

Swanson, D. A., Tayman, J. & Bryan, T. M. (2011). MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts. *Journal of Population Research*, 28, 225-243.

Stonehouse, J. & Forrester, G. (1998). Robustness of the *t* and U tests under combined assumption violations. *Journal of Applied Statistics*, 25(1), 63 – 74.

Tayman, J. & Swanson, D. A. (1999). On the Validity of MAPE as a Measure of Population Forecast Accuracy. *Population Research and Policy Review*, 18(4), 299-322.

Toda, M. (1963). Measurement of subjective probability distributions. ESD-TDR-63-407, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, Bedford, MA.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458.

Verdolini, E., Anadon, L. D., Lu, J. & Nemet, G. F. (2015). The effects of expert selection, elicitation design and R&D assumptions on experts' estimates of the future costs of photovoltaics. Energy Policy, 80, 233-243.

Visser, P. S., Krosnick, J. A., Lavrakas, P. J. & Kim, N. (2014). Survey research. In Reis, H. T. & Judd, C. M. (Eds.), Handbook of research methods in social and personality psychology (Second Edition). New York: Cambridge University Press.

Wackerly, Mendenhall & Scheaffer (1996). Mathematical Statistics with Applications (Fifth Edition). California, USA: Duxbury Press.

Wilk, M. B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1), 1 -17.

US Army Corps of Engineers. Corps Risk Analysis Gateway. Expert Elicitation. (October 6th, 2017). Retrieved from: http://www.corpsriskanalysisgateway.us/lms/course.cfml?crs=15&crspg=233

Usher, W. & Strachan, N. (2013). An expert elicitation of climate, energy and economic uncertainties. *Energy Policy*, 61, 811-821.