AFRL-RH-WP-TR-2020-0139

INTERPRETING GRAPHS IN A WISDOM OF CROWDS FORECASTING INTERFACE FOR INGELLIGENCE ANALYSTS

Mary E. Frame, Ph.D.
Anna Maresca
Wright State Applied Research Corporation
4035 Colonel Glenn Hwy.
Beavercreek, OH 45431

Michaela Schwing
Bradley Schlessman, Ph.D.
711th HPW/RHWA
Warfighter Interactions & Readiness Division
2255 H. Street
Wright-Patterson AFB, OH 45433

OCTOBER 2020

INTERIM REPORT

AIR FORCE RESEARCH LABORATORY
711th HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WARFIGHTER INTERACTIONS & READINESS DIVISION
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the AFRL Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2020-0139 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
BRADLEY R. SCHLESSMAN, DR-III, Ph.D.
Work Unit Manager
Mission Analytics Branch
Warfighter Interactions & Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//
WILLIAM P. MURDOCK, DR-IV, Ph.D.
Chief, Mission Analytics Branch
Warfighter Interactions & Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//
ALFREDO RIVERA, Col, USAF
Acting Chief, Warfighter Interactions & Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-10-2020	Interim	14 February 2020–14 November 2020

4. TITLE AND SUBTITLE
Interpreting Graphs in a Wisdom of Crowds Forecasting Interface for Intelligence Analysts

5a. CONTRACT NUMBER
FA8650-19-F-6052

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Mary E. Frame, Ph.D.* Anna Maresca*
Michaela Schwing** Bradley Schlessman**

5d. PROJECT NUMBER

5e. TASK NUMBER
0007

5f. WORK UNIT NUMBER
H0Y3

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

*Wright State Applied Research Corporation
4035 Colonel Glenn Hwy.
Beavercreek, OH 45431

** 711th HPW/RHWA
Warfighter Interactions & Readiness Division
2255 H. Street
Wright-Patterson AFB, OH 45433

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Airman Systems Directorate
Warfighter Interactions & Readiness Division
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHWA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-WP-TR-2020-0139

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
Report contains color. AFRL-2021-2234; Cleared 26 Jul 2021

14. ABSTRACT

Intelligence analysts must integrate highly dynamic information in order to solve problems reactively and make predictions about future events proactively. Individual predictions may or may not be highly accurate, depending on the expertise of the forecaster. However, previous research on the wisdom of crowds has determined that often, aggregated estimates of multiple experts is even closer to the truth than most single expert forecasts. We have developed a tool as a means of aggregating and displaying the wisdom of crowds for Intelligence, Surveillance, and Reconnaissance (ISR) analysts. This tool allows analyst forecasters to make their own predictions regarding self-selected questions pertinent to mission objectives. Additionally, the tool displays the aggregate predictions of other analysts over a 1-month period, along with an individual analyst's reasoning for each of their predictions. This allows any given analyst to consult the wisdom of the crowd and the opportunity to update their predictions, if desired. In order to ensure this tool is maximally effective and interpreted properly by analysts, it is critical that the graphical display of the crowd's prediction information is comprehensible. We tested a variety of graph types (two forms of line graph, a bar graph, a box plot, and a dot plot) to display the wisdom of crowds with varying features to ascertain individuals' ability to accurately understand the presented information using a small pilot sample of participants

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bradley Schlessman, Ph.D.
U	U	U	SAR	24	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

TABLE OF CONTENTS

LIST OF FIGURES ii
1.0 SUMMARY 1
2.0 BACKGROUND ON FORECASTING IN INTELLIGENCE ANALYSIS 2
2.1 Forecasting 2
2.2 Wisdom of Crowds (WOC) 4
2.3 How People Interpret Information in Graphs 5
3.0 CURRENT RESEARCH EFFORT: THE SPHINX FORECASTING TOOL 7
4.0 METHOD 9
4.1 Participants 9
4.2 Graphs 9
4.3 Procedure 11
5.0 RESULTS OF INITIAL STUDY 12
6.0 DISCUSSION OF PILOT STUDY 15
6.1 Implications of the Study 15
6.2 Future Directions 15
7.0 REFERENCES 17
8.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS 19

LIST OF FIGURES

Figure 1. Example of Graph and Question Display from Galesic & Garcia-Retamero (2011) Involving Medical Information. 8

Figure 2. Example of each Graph Type with N = 15 Data Points. 10

Figure 3. Accuracy Results from the Initial Pilot Study. 12

Figure 4. Graph Type and Split of Crowd Forecast 13

Figure 5. Graph Type and Degree of Change 14

1.0 SUMMARY

Intelligence analysts must integrate highly dynamic information in order to solve problems reactively and make predictions about future events proactively. Individual predictions may or may not be highly accurate, depending on the expertise of the forecaster. However, previous research on the wisdom of crowds has determined that often, aggregated estimates of multiple experts is even closer to the truth than most single expert forecasts. We have developed a tool as a means of aggregating and displaying the wisdom of crowds for Intelligence, Surveillance, and Reconnaissance (ISR) analysts. This tool allows analyst forecasters to make their own predictions regarding self-selected questions pertinent to mission objectives. Additionally, the tool displays the aggregate predictions of other analysts over a one-month period, along with an individual analyst's reasoning for each of their predictions. This allows any given analyst to consult the wisdom of the crowd and the opportunity to update their predictions, if desired. In order to ensure this tool is maximally effective and interpreted properly by analysts, it is critical that the graphical display of the crowd's prediction information is comprehensible. We tested a variety of graph types (two forms of line graph, a bar graph, a box plot, and a dot plot) to display the wisdom of crowds with varying features to ascertain individuals' ability to accurately understand the presented information using a small pilot sample of participants.

2.0 BACKGROUND ON FORECASTING IN INTELLIGENCE ANALYSIS

Intelligence analysts must make inferences and predictions by aggregating diverse, highly dynamic information that updates continuously each day. Information to solve needed problems may come from a variety of sources, leading to different individuals making different predictions about the likelihood of future events, depending on what information is readily available to them. However, often in the aggregate, there may be a strong wisdom discerned by a crowd of relevant experts, which could and should be leveraged by others when making predictions under uncertainty. Forecasting tournaments, such as the Good Judgment Open (GJO) project (Ungar, et al., 2012) and interfaces that provide insight into the predictions of others may assist novices or less experienced experts to leverage information provided to them more effectively. However, in order for these tools to be maximally efficacious, it is critical that the collective opinions of the crowd are displayed in a manner that can easily be interpreted and utilized effectively.

Furthermore, it is valuable for analysts to observe how individual and crowd forecasts have been modified over time as events in the world are continuous, and critical information may lead to fluctuations in aggregate opinions.

2.1 Forecasting

As one Central Intelligence Agency (CIA) intelligence officer cited in a recent op-ed (Bobby, 2019), forecasting is difficult within the intelligence community. Errors are often attributed to cognitive biases, groupthink, faulty mindsets, stove-piping of information, individual overconfidence, confirmation bias, anchoring, and outdated frameworks (Bobby, 2019; Wintle, et al., 2012). More broadly, humans have shown to have difficulty making forecasts even when there is sufficient information and the individual has sufficient cognitive capacity to make the prediction. Among experts, this may be due to a high degree of proficiency with making linear extrapolations to known past data, which are often accurate for future predictions (Bobby, 2019). However, this can lead to analyst overconfidence and an overreliance on linear trend expectations, which may lead to faulty decision making in cases where a linear extrapolation (i.e. past behavior is indicative of trends for future behavior) is inappropriate. Furthermore, the longer complex systems have to mutate and evolve, the more stochastic the behavior of that system becomes, making its future behavior even harder to predict. This is why it is critical for visualizations to provide information on both the general trend of data, but also the degree of stochasticity and uncertainty. Certain types of graphical displays can enhance this interpretability, but other visualizations can make this information difficult to glean. This underscores the value of testing multiple graphical display types to ensure they are conveying the correct trend information.

When looking at forecasts in groups, certain factors of individuals and teams improve forecast prediction and decision-making. Forecasts are typically more accurate when they are made within a group of diverse individuals (Bobby, 2019; Wintle, et al., 2012). However, certain skills and traits have been identified as traits of good analysts and forecasters including: high cognitive ability, political knowledge, strong inductive reasoning/pattern detection, and high fluid intelligence (Bobby, 2019). Furthermore, on a forecasting team, traits including: balance between collaboration and competitiveness, an open-minded approach to problem solving, and a commitment to self-improvement were seen as important attributes. Regardless of personality or skills, providing forecasters with training, tracking of past events, and teaming can greatly reduce the error in a forecast (Mellers, et al., 2014).

Forecasting is defined as predicting the probability that a specific event will occur within a certain timeframe; these predictions provide situational awareness (Ramakrishnan, et al., 2015). Two of the most common types of forecasting are point predictions and range forecasts (Haran & Moore, 2014). Point predictions are used when a forecaster is trying to determine what event, specifically, will occur. The downside to this type of forecast is that it can highlight the forecaster's overconfidence in their answer, making them reticent to voice any uncertainty. On the other hand, range forecasts allow the forecaster a margin of error in their estimates. Although a range does allow forecasters to voice their uncertainty, it is not perfect, as precise estimates contain lower confidence and have a narrow range of values and conservative estimates contain higher confidence and a much wider range of values (Haran & Moore, 2014). Although forecasts can be made by anyone based on available information, this skill is particularly critical for intelligence analysts. In addition to closed intelligence information (Human Intelligence [HUMINT], Signals Intelligence [SIGINT], etc.), open-source tools can be utilized by analysts to collect information for making an accurate forecast, including: social media (Facebook, Twitter), physical measurements (temperature, humidity), or unconventional surrogates (reservations at hotels, security footage from a civilian building).

Particularly within military operations, there is a strong desire to develop intelligence analyst skills to become better forecasters. Improved forecasting capability is desirable so that analysts can make better decisions and effectively leverage information provided by other expert analysts' forecasts. Justifications and explanations of the rationale behind a forecast can be particularly valuable for aiding forecasting by analyst colleagues. An Intelligence Advanced Research Projects Activity (IARPA) program called Aggregative Contingent Estimation (ACE) is working to determine how we can best identify the knowledge and understanding of an expert forecaster. This would give greater weight to their estimates compared to novices (Bobby, 2019; Forlines, et al., 2012). Forecasts are often measured using the Brier score (Brier, 1950). The Brier score is applicable when there are discrete, binary responses rated using a probability percentage (Brier, 1950; Rufibach, 2010). This is generally the preferred metric for measuring forecasting aptitude rather than binary accuracy (correct/incorrect percentages) because it incorporates the certainty magnitude of the response. In a scenario, for example, where an event does occur, the Brier score incentivizes correct predictions that are strong (e.g. 95% likely to happen) rather than hedging (51% likely to happen) and is more penalizing of incorrect predictions that are strong but wrong (e.g. 5% likely to happen) than hedging (e.g. 49% likely to happen). The score can only be computed once the event window has closed, with lower Brier scores indicating higher performance on prediction (Forlines, et al., 2012). Brier scores are calculated using the following formula:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (1)$$

Where f_t is the probability of the forecast, o_t is the binary outcome of the event at time t (0 or 1), and N is the number of individual forecasts. The individual's score is cumulative over time, meaning the Brier score for each participant is updated at the end of each month.

2.2 Wisdom of Crowds (WOC)

Previous research has established that, under certain circumstances, judgments or estimates may be more accurate in the aggregate within a group than most individual judgments alone, due to the averaging over of individual estimators' biases. This collective accuracy is referred to as the wisdom of crowds. One of the earliest and most well-known examples illustrating the wisdom of crowds comes from Galton (1907). He conducted an experiment at a local gathering during which he asked the public to guess the weight of a dressed ox; roughly 800 participants paid an entry fee and then privately made their guess in pounds. At the conclusion of the day, Galton aggregated all of the guesses and found that the collective crowd was within one percent of the actual weight.

There are several different kinds of inputs the crowd can offer including votes, preferences, probability, quality, relevance or rank, and numerical value. Members of a group can also provide a response along with a probability estimate that represents their confidence in their own answer. This gives people who are more confident in their answers more weight in the overall outcome and is especially used in forecasting, see weighted averaging formula below:

$$WLA = \frac{1}{N}\sum_{i=1}^{N} w_i j_i \tag{2}$$

There are multiple means of measuring the wisdom of a given crowd for a third-party observer. The simplest way is to measure the single best individual, or the person who is perceived as having the most knowledge and expertise (Mannes, et al., 2014). This approach is favored when there is a high degree of expertise variability in the group and there is a valid cue for identifying the "best member." for example, requesting the assistance of a medical expert on a plane, rather than taking a poll of a subset of the passengers. A select crowd approach is to subsample a small group of the full crowd that might have more knowledge than an average member, but whose expertise cannot easily be differentiated. The size of the select crowd is dependent on the overall size of the whole crowd, however a good rule of thumb is to use the top five members of the whole crowd. Finally, there is the whole crowd approach, which as it sounds involves taking the votes or choices of the entire crowd into consideration. This approach is optimal when there is little variability in expertise between crowd members and members tend to not show a particular bias on one side of the truth.

Certain factors contribute to the success of wisdom of the crowd, including group size, diversity, and expertise. When the group is sufficiently large enough, aggregation to the median or mean typically provides a response that is more suited to the truth, due to cancelling out sources of variable noise (Bennett, et al., 2018). Fortunately, these simple measures of central tendency have been found to be just as accurate as more sophisticated aggregation methods (Mannes, et al., 2014). In situations where the group must make an outright judgment, it is important to include as many members as possible to get a large enough sample of the population (Lyon & Pacuit, 2013). However, Galesic & Garcia-Retamero (2011) found that moderate sized groups, as few as 14 members (Bachrach, et al., 2012), can perform better than individuals and large groups when the task at hand is relatively easy.

Expertise of the crowd also contributes to higher accuracy in the wisdom of crowds (Iyer & Graham, 2012). This need for expertise can be mitigated with crowd diversity, meaning that

not all members have the same expertise. It is not necessary for all members of the crowd to have the knowledge to give a perfectly accurate response, as the members who do not possess this knowledge provide the diversity in the group. Maintaining this diversity can be difficult when there is cross-communication among crowd members. It is important to avoid situations that can cause social influence or group think to reduce the crowd's diversity, as this can also reduce the efficacy of the group (Lorenz, et al., 2011). The herding effect, when a large portion of the group thinks in the same way regardless of accuracy, can also dramatically change the outcome of the wisdom of a crowd; there can be several herds within the same group. Another critical aspect of the original Galton (1907) study is that members of the group were motivated by opting in and paying a fee to guess. Bennett et al. (2018) found that when members of a group decide to opt-in and provide a response freely, the aggregate wisdom of the crowd is more accurate than when members are compelled to respond.

2.3 How People Interpret Information in Graphs

As valuable as the wisdom of crowds is for improving decision making, it cannot be leveraged as an effective decision aid unless it is presented such that an average viewer can accurately interpret the perspective of the crowd. Individual and aggregate quantitative data can be displayed using one or multiple graphs (Friel, et al., 2001). It is critical for an individual to accurately interpret graph information for effective decision-making and reasoning. Graphical displays therefore should communicate the most important elements that will need to be integrated by a decision-maker (Friel, et al., 2001; Okan, et al., 2012). Effective information communication in graphs can additionally help overcome biases (Walker, et al., 2019). For forecasting decisions, this information should be easily gleaned and interpreted, due to time pressure. Previous research on graph interpretation provide insights into what design features on graphs correlate with cognitive processing and provide a baseline for testing multiple potential display types (Friel, et al., 2001). Furthermore, there appear to be measurable individual differences in graph comprehension and graph literacy (Galesic, & Garcia-Retamero, 2011).

Shah and colleagues (1999) found that graph displays that visually chunk information that the observer needs to compare better support perceptual processes and aid in graph interpretation by reducing the need for complex inferential processes. They found that when temporal information about multiple geographic regions was displayed using a bar chart, students were able to draw factual inferences about the geographic regions in comparison to one another, but did a poorer job at understanding how each region experienced comparative changes over time. However, since the graphs were being used to demonstrate changes in each region over time, this led to a lack of appropriate comprehension for most students. By modifying the bar charts into line plots with identical information, as well as splitting the plot into two plots with fewer variables and transforming the y-axis into a percentage value, this emphasized the necessary temporal relationship for how each region changed over time and improved student comprehension. The trade-off between understanding factual vs temporal information illustrates that which graph is appropriate to utilize is dependent on the information the observer needs to glean. The Gestalt principles of proximity, similarity, connectedness, and continuity are used to determine how people visually chunk information presented in graph form (Peebles & Ali, 2015). Trends over time are easier to identify when using line plots due to connectedness and continuity, while comparisons within a given time frame are easier to identify when using bar charts due to proximity and similarity. In other

words, differences within clusters of bars on a bar chart are easier to identify than differences across clusters. However, the perceptual organization of data into visual chunks is more likely to influence users than the format of the chart.

Graph literacy is a trait that varies between individuals. Individuals with high graph literacy tend to extract more complex information from line graphs and can better infer the main effects from bar graph displays (Okan, et al., 2012; Shah & Freedman, 2011). Galesic and Garcia-Retamero (2011) created a methodology for testing multiple types of graphs and levels of interpretation of the information within those graphs. In their first level, *reading the data,* means that an individual should be able to glean direct information from a graph, for example, reading the height of a bar on the y-axis. At the second level, *reading between the data*, individuals need to glean relationship information from the data, such as mentally calculating the difference between two bars or summing the values of multiple bars. Finally, the third level of *reading beyond the data* involves making inferences and predictions based on the data, such as forecasting a trendline or noting seasonal effects.

This methodology is similar to the three levels of comprehension discussed by Friel that include translation, interpretation, and extrapolation/interpolation. Translation focuses on gleaning specific data from a graph. Interpretation involved identifying relationships between the variables. The extrapolating level involves making predictions based on the relationships discovered within the data (Friel, et al., 2001). Certain graph types may be more suited for conveying each of these levels interpretation for wisdom of crowds information, and both individual level and aggregate crowd data may be necessary to answer particular questions. As such, multiple graph types will be tested based on their proposed strengths and weaknesses. For example, line graphs are particularly suited for displaying temporal or functional relationships (Friel, et al., 2001). Violin plots provide the same quantile and mean information as a typical box plots, but with the added benefit of displaying the distribution of the data. Bar and pie charts can both be utilized to display quantities and percentages, respectively. However, bar plots are typically recommended over pie charts since most people make more accurate judgments of differences on a common scale than they do of angle differences (Simkin & Hastie, 1987). Finally, dot plots allow individual data to be visualized similarly to a histogram where density of dots reflects a greater number of individuals who responded in a like manner (Friel, et al., 2001).

3.0 CURRENT RESEARCH EFFORT: THE SPHINX FORECASTING TOOL

The Sphinx forecasting tool was developed to display the wisdom of crowds of expert intelligence analysts for the purpose of improving analyst forecasting ability. By displaying prediction information aggregated across dozens of intelligence analysts, the goal is to improve the overall prediction ability of each analyst. Bennett and colleagues (2018) found that allowing individuals to opt-in to questions of expertise was beneficial to the overall wisdom of the crowd, creating an exceptional crowd who were motivated to be as accurate as possible due to paying for the privilege. Similarly, all of the forecasting questions are provided by the Sphinx tool to analysts and they must opt-in to respond. Importantly, this task extends beyond the analysis performed by Bennett and colleagues (2018) as forecasting requires making inferences beyond the current data set, in line with the third level of the graph theory proposed by Galesic & Garcia-Retamero (2011). They can choose to respond immediately if they have some expertise on the topic, or they can wait until others respond to assess the wisdom of the crowd before making their forecast. Likewise, if they feel unable to make a prediction even with the wisdom of the crowd, they are permitted to not answer questions.

Analysts are given a group of time-sensitive questions about events that may or may not occur at the end of the month. They individually provide their forecast as a probability percentage of whether the event will occur. This allows the prediction to not only provide a "yes/no" binary response to whether an event will occur or not, but also provides a rating of confidence. As a non-sensitive example, forecasters could be asked, "What is the probability that Simone Biles will win the 2020 Olympics all-around gold medal?". Forecasters would be afforded the opportunity to view a display of the distribution of other raters in addition to their justifications for their responses. Importantly, if this visualized crowd information is unclear or confusing, this can lead to misinterpretation and actually decrease the accuracy of an individual's forecast.

Although not consequential for this non-sensitive question example, in real-world intelligence analysis, the consequences of systematic errors of interpretation can be dire.

Within the Sphinx tool, each forecaster may change his or her prediction over time. For example, given Biles' record, the crowd might generally fall above 50% (a prediction that she will win gold), but if she were to suffer an injury or be discussing retirement, individuals could change their answers and move the needle for the crowd. In addition to providing responses to individual questions, as time progresses, monthly Brier scores are tabulated for each individual based on their accuracy and confidence of response. This is valuable from a research prospective, as this affords us the opportunity to see who changes predictions (i.e,. high versus low expertise), how often they make changes, and why they make changes, via their written justifications with each submission of a new answer.

In order to provide maximum efficacy of the wisdom of crowd data, it is important that it is summarized in graphs that can be adequately interpreted by experts and novices. That is to say, it is important that a graphical display be capable of communicating information that allows even someone with less expertise to make Level 2 (relational) and Level 3 (extrapolation) interpretations about the perspective of the crowd. If graphs are presented in such a manner that a typical user will misinterpret the conclusion displayed in the graph, this will lead to less effective information integration based on the wisdom of crowds. This may further mislead the group into suboptimal decision-making, rather than providing an improvement. We are interested in understanding which types of graphical displays will allow for forecasters to make accurate

assessments about the wisdom of the crowd based on Galesic's (2011) levels of graph interpretation: direct interpretation (what does the crowd say), relational interpretation (how does the crowd opinion change over time), and inferences beyond the data (what will the crowd perspective be at some time in the future). Figure 1 provides an illustration of questions from each level from their 2011 paper.

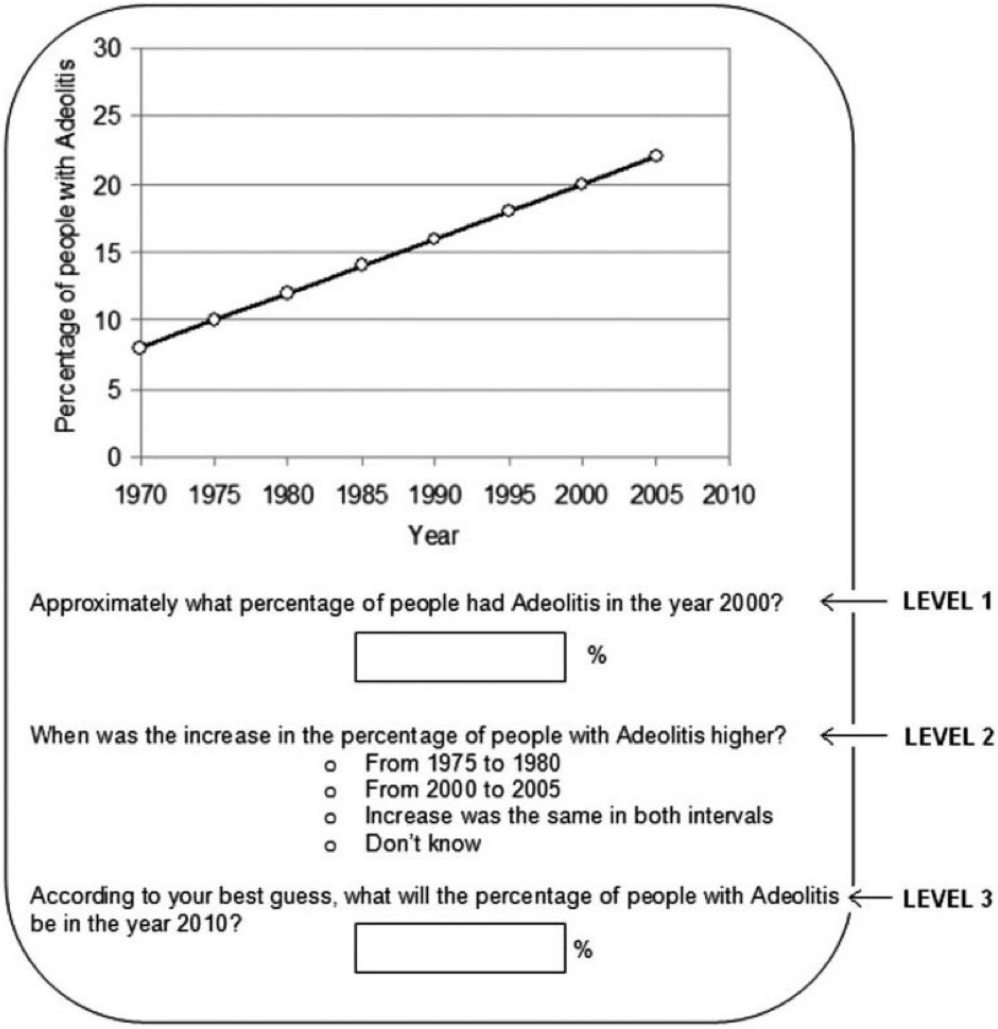

Figure 1. Example of Graph and Question Display from Galesic & Garcia-Retamero (2011) Involving Medical Information.

*Stimuli were presented similarly in the present experiment, with a graph displayed with a single question at a particular level of complexity. Each graph was presented with all 3 levels of question in sequential order***.**

4.0 METHOD

4.1 Participants

For the initial small pilot study, we studied a group of seven subjects who were not expert intelligence analysts. They were slightly familiar with the concept of a forecasting tournament such as GJO. This was done to calibrate the types of graphs chosen and to collect a larger sample of questions than would be possible with a full expert subject pool.

4.2 Graphs

Graphs for the study were generated using Microsoft Excel. Each graph was static and presented within a mock presentation of the Sphinx tool interface. We generated mock wisdom of crowd data for questions from four forecasting topics: geopolitical, financial, pop culture and sports.

This was done to prevent expertise bias. Each data set was designed to have been collected over the course of three months, where a number of forecasters could make and update their predictions. As the question content is not as critical as the interpretation of the crowd's prediction, we did not tailor the presented questions based on subject-matter expertise, nor did we require participants to search for this information. We did not present any of the individual prediction justifications, only the aggregate forecast graphs. This differs slightly from the Sphinx tool interface, where graphs are somewhat interactive, allowing participants to narrow the displayed date range, up to a maximum view of a full month.

Graphs were blocked using a full factorial of: size of data set (2), degree of split for yes vs no responses (2), degree of change over time (2), and graph style (5), for 40 trials in total. Landwehr and Watkins (1986) found that the size of the data set influences the appropriateness of different graph styles. We varied the number of forecasters' predictions represented on the graph to explore if there is a trade-off in which graph type is most interpretable and appropriate to use based on a large (200) versus small (15) number of respondents. This small group size is based on the findings of Bachrach et al. (2012), who found that approximately 14 members was a sufficient crowd for efficacious prediction. The second factor was the degree of response split. In one condition, the crowd was split nearly 50/50 between a "yes" and "no" prediction, and in the second condition, the majority of the crowd favored one response over another, with a 75% favorability toward either the "yes" or "no" prediction. The third manipulation was the degree of change over time in the predictions. Half of the graphs were relatively stable with only small (up to 5% fluctuations) and the other half had larger fluctuations in preference (up to 15%). Finally, each condition was displayed using five different styles of graph, varying type based on the proposed strengths and weaknesses of graphs delineated by Friel and colleagues (2001). The graph styles chosen were: (1) line plot with group average, maximum, and minimum, (2) line plot with proportion of respondents above vs. below 50%, (3) stacked bar plot, (4) box-and-whisker plot, and (5) dot plot. Figure 2 illustrates an example of each of these 5 styles of graph with three months of aggregated crowd data.

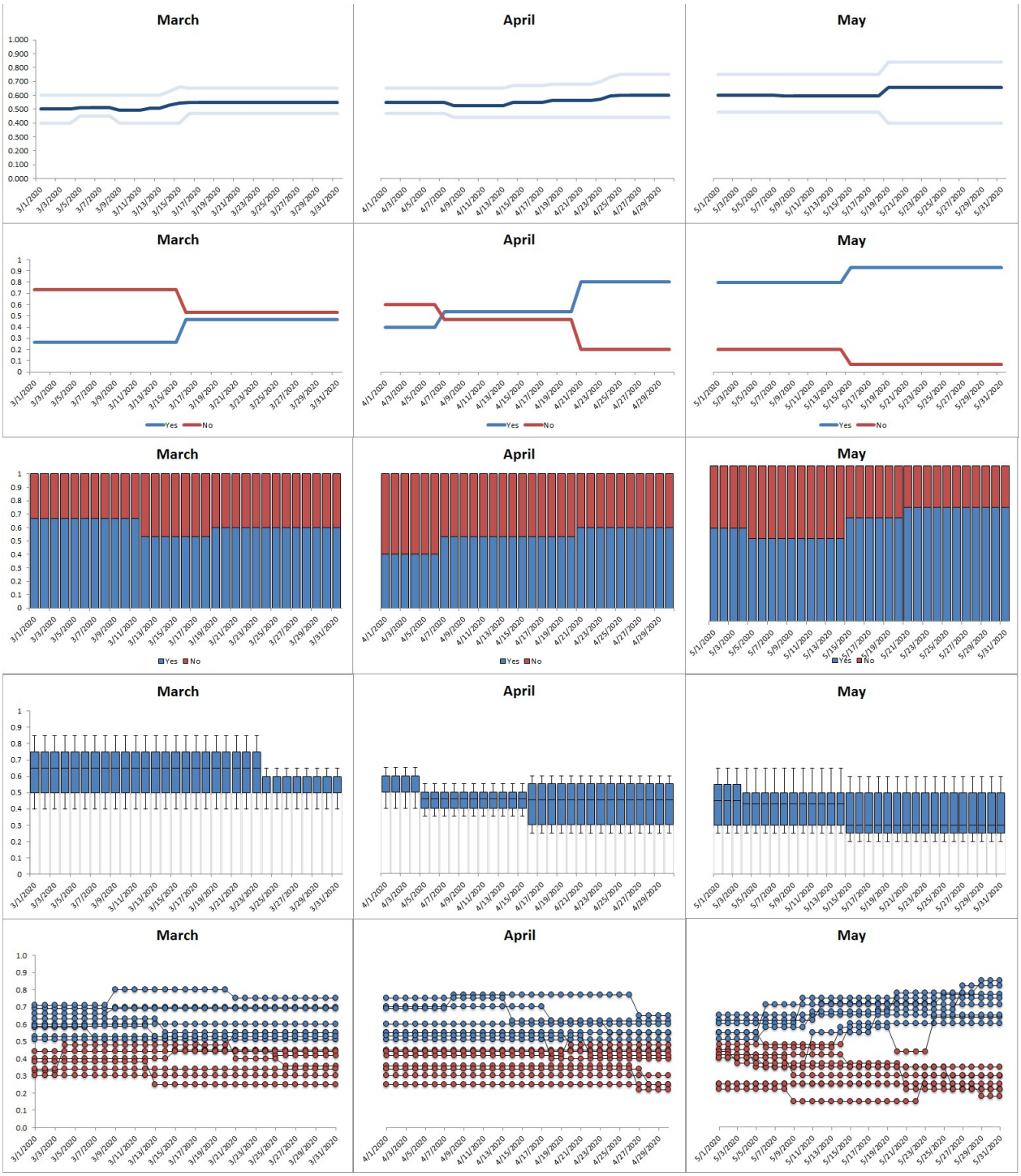

Figure 2. Example of each Graph Type with N = 15 Data Points.

Top to bottom: Line Graph (mean, maximum, minimum), Line Graph (proportion above/below 50%), Bar Graph (proportion of crowd above/below 50%), Box Plot, and Dot Plot. "Yes" and "No" are coded as "will happen" and "won't happen" respectively based on whether an individual's forecasting estimate was above or below 50%.

4.3 Procedure

The wisdom of crowds forecasting graphs were programmed in Qualtrics, and each participant was emailed a participation link. The experiment took around 80 minutes to complete.

Following an informed consent procedure and instructions, participants would see three months of graphed forecasting data. Under each graph were three questions of increasing complexity, similar to the questions from Galesic and Garcia-Retamero (2011):

- Level 1 – Understanding the material in the graph "what does the crowd say?"
- Level 2 – Understanding the relationships presented in the graph "how does the crowd opinion change over time?"
- Level 3 – Predicting beyond the data in the graph "What will the crowd perspective be at a future time if the trend in the data continues?"

Responses to Level 1 questions were made using a categorical multiple-choice response with percentage values in ranges of 10% (e.g., 0-10% probability, 11-20% probability, etc.). An example of a Level 1 question might be "What percentage of the crowd believed that the spot price of silver would be over $17.50 per ounce on June 1?" For Level 2 and Level 3 questions, responses were chosen from a multiple-choice selection. A Level 2 question would be "Which day in March represented an inflection point where the crowd became more confident that the US women's team would win the World Cup?" Level 2 is also an appropriate level for anomaly detection, for example, in a sports related question, one might ask which day appears to be indicative that the sports team has had a series of injuries. Finally, a Level 3 question example would be "Based on your best guess, what will the opinion of the crowd be on July 1 regarding the probability of Russia declaring a national health emergency?" Both levels 2 and 3 could be classified as extrapolation and interpolation by Friel and colleagues (2001) whereas Level 1 requires only interpretation under their framework.

For each graph type, there were three additional questions, for a total of 15, randomized across conditions. There was an attention check question to ensure reasonable responding, a Level 3 question that asked the same question but with a different hypothetical crowd, and a Level 3 question that asked a similar question, but with the same hypothetical crowd. The 3 Level questions from Galesic & Garcia-Retamero (2011) allow us to determine if there are differences in graph literacy based on the manipulated factors, and the additional questions allow us to determine how capable the crowd is at extrapolating beyond the data based on knowledge. At the end of the graph interpretation block, participants were asked a series of demographic questions, questions about their proficiency in each of the question categories, as well as which graphs they found easiest and most difficult to use.

5.0 RESULTS OF INITIAL STUDY

For the initial pilot testing all five graph types, we collected data from seven individuals with a mean age of 37. Most participants reported having an advanced degree (master's degree or Ph.D.) but did not self-report as being particularly adept with graphs ($M = 3$, from a 5-point Likert scale).

On average, participants self-reported that they had low knowledge on all categories (finance, sports, and pop culture) based on a 5-point Likert scale ($M = 2$) except for on geopolitical which had a slightly higher mean familiarity score of 2.4. The majority of the sample reported that the bar graph was easiest to interpret, with a plurality preferring the Yes/No line graph. The dot plot was reported by all members of the sample to be the most difficult to interpret chart.

We examined accuracy based on the level of difficulty of the question. There were no significant differences in mean accuracy based on the level of the question (all ranged between 57% and 61%). The scores were nominally lower for the Level 3 questions that required extrapolation to a different crowd ($M = 49\%$) or extrapolation to a different question ($M = 51\%$), but these were not significantly different from the main 3 levels of question difficulty (see Figure 3). There was a significant interaction however between accuracy by difficulty level based on the type of graph that was interpreted, $F(4,34) = 11.34$, $p < .01$. For the Level 1 (direct interpretation) questions, accuracy was significantly higher for both the Yes/No Line and Bar charts than the Dot plot and Mean Line plot. For the Level 2 (relationship between multiple times) questions, accuracy was significantly higher both the Yes/No Line and Bar charts than the Box and Mean Line plots. For the primary Level 3 (extrapolating beyond the data) questions, accuracy was higher for the Yes/No line graphs compared to the Dot plots. The right panel of Figure 3 illustrates the results for the specific Level 3 questions. For the Different Crowd questions, accuracy was significantly higher for the Bar and Yes/No line charts than the Box and Dot plots. For the Different Question prompts, accuracy was significantly higher for the Bar plot compared to the Mean line and Yes/No line plots. This represented a reversal of the dominance of the Yes/No line plot.

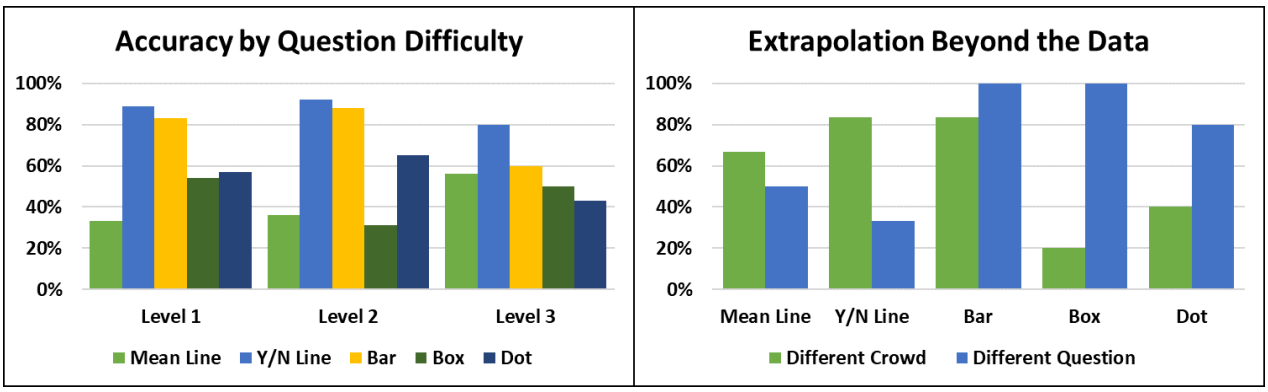

Figure 3. Accuracy Results from the Initial Pilot Study.

For the primary questions at all three levels, highest performance was achieved with the Yes/No line graphs and the Bar charts. For the questions that required inference to a different crowd, the Yes/No Line and Bar plots yielded the best performance, but for the different question extrapolation, the highest performance was yielded by the Bar and Box plot visualizations.

Performance significantly varied as a function of graph type when crossed with question difficulty. There was a significant main effect of graph type $F(4,174) = 1.78$, $p = .04$ and a significant interaction, $F(4,174) = 8.38$, $p < .01$. For Level 1 and Level 2 questions & different crowd extrapolation, Bar and Y/N line are equally highest, whereas for Level 3 questions, Y/N line yielded highest performance. Regarding extrapolation questions, extrapolating to a different question yielded significantly higher performance with the bar chart. Accuracy was generally higher for easier (Level 1) questions compared to harder questions (Level 3), in the predicted direction.

In addition to questioning difficulty, we were interested in examining various potential forecasting crowd factors on visualization interpretability and accuracy. Accuracy was examined based on graph visualization type and degree of response split between will happen/won't happen. There was a significant main effect of graph type, $F(4,69) = 16.96$, $p < .01$ and a significant interaction, $F(4,69) = 2.97$, $p = .02$ (see Figure 4). For most of the graph types, there was no significant difference based on the proportion of yes vs no predictions. However, there was an interaction with the dot plot, where the stronger split of the crowd yielded significantly higher accuracy. The highest overall accuracy was yielded by the bar chart and yes/no line visualizations.

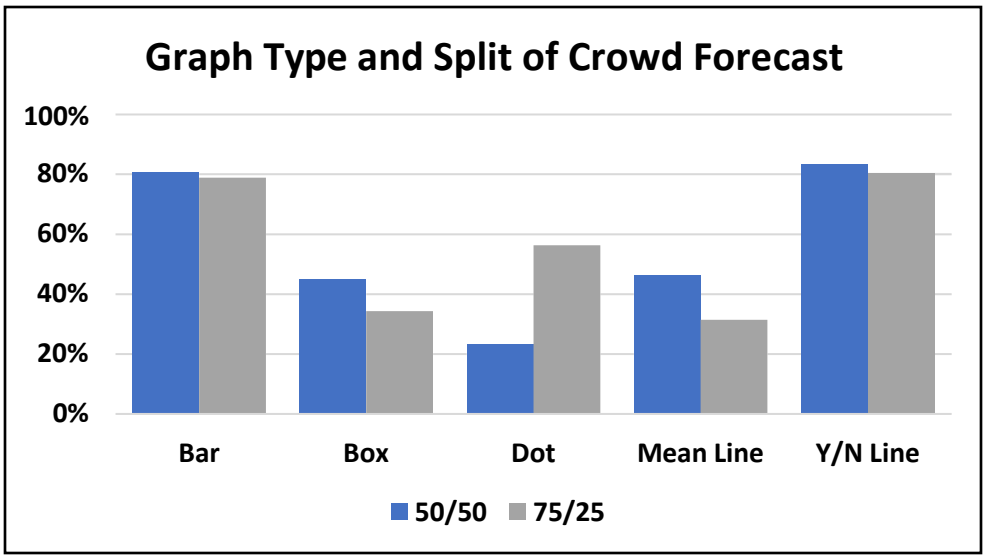

Figure 4. Graph Type and Split of Crowd Forecast

Generally, there were no significant differences in participant accuracy for each visualization based on the degree of crowd split, but the general trend was that performance was higher when there was closer to a 50/50 split between "will happen" and "won't happen" with the crowd. However, this pattern significantly differed in the opposite direction for the dot plot visualization, with the 75/25 crowd split yielding higher accuracy.

The final two factors examined were graph type and the degree of crowd opinion change over time. For half of the trials, graphs portrayed a change of approximately 5% whereas the other half portrayed a crowd fluctuation of approximately 15%. There was no significant interaction between these two factors, but there was a significant main effect of graph type $F(4, 69) = 14.06$, $p < .01$. Again, the highest performance was seen with the Bar charts and the Yes/No line plots (see Figure 5).

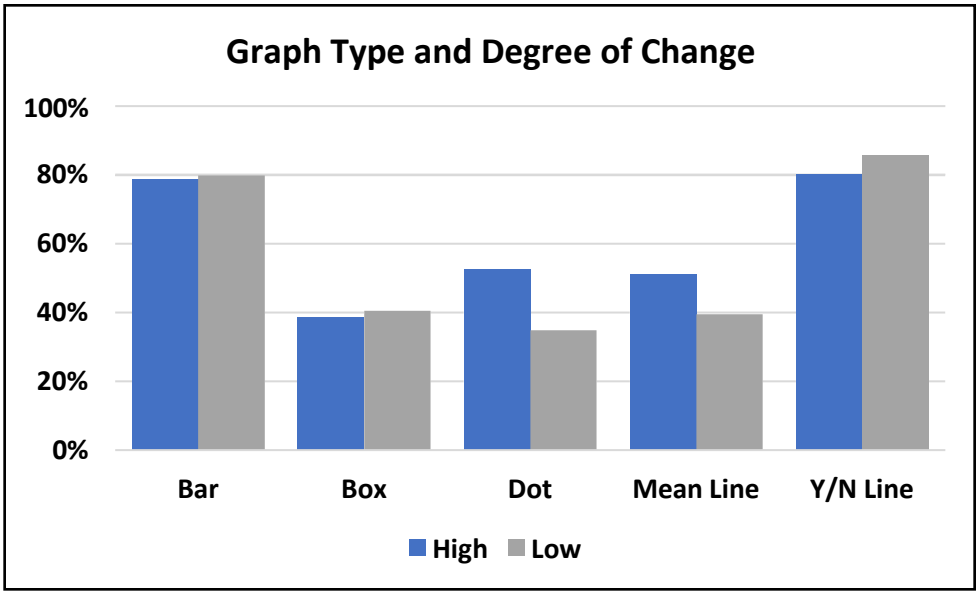

Figure 5. Graph Type and Degree of Change

There were no significant differences in performance in interpreting visualization based on whether there was a high or low change in mean crowd prediction over time.

6.0 DISCUSSION OF PILOT STUDY

6.1 Implications of the Study

The primary focus of this study was to determine if there were differences in participants' ability to understand different graph visualizations when making inferences about crowd forecasts. We further examined if visualizations facilitated understanding under varying crowd factors: degree of split between Yes/No responses, and the degree of change in the crowd's forecast over time. The outcome provides prescriptive recommendations for the Sphinx tool. Regardless of the difficulty of the question, the line plot and bar charts that illustrated the proportion of "yes/will happen" responses and "no/will not happen" responses facilitated consistently high performance. For both of these charts, there were no significant effects of crowd split or degree of crowd opinion change, so the plots are robust to variable crowd data. The bar plot was particularly faciliatory for the Level 3 responses, as accuracy on both generalizing to different crowd compositions and different questions were highest for the bar plot visualization. Thus, if a particular Sphinx forecasting question requires forecasters to generalize beyond the given crowd data, a bar plot may be preferable to a line plot.

Self-reported preferences for the graphs did seem to strongly correlate with performance. Participants self-reported that they preferred the bar plots or Y/N line plots the most and that they preferred the dot plots the least. Although performance when interpreting dot plots was not particularly low, performance was highest for the Y/N line plots and the bar charts. There was no correlation between performance and familiarity with interpreting graphs. This indicates that high performance in interpreting the crowd's forecasts did not rely on being particularly adept with understanding complex graphs. Indeed, each of the chosen graph types was designed to be relatively simple, with a limited amount of data conveyed on each one. None of the graphs displayed had multiple axes or multiple data sources.

6.2 Future Directions

The Sphinx WOC Forecasting tool is still in development, and thus visualizations are continuing to be added. Although this pilot study has helped to solidify the graphical displays on the individual question level, there is an interest in presenting data hierarchically for the tool. For example, it may be of interest to aggregate questions by domain (e.g., Sports, Geopolitical), or base questions on a on subsample of the forecasters based on expertise. In the full version of the Sphinx tool, questions are organized based on intelligence Priority Information Requirements (PIR) with each prediction question representing various facets to answer the PIRs. At the summary level, the current plan is to provide visualizations using pie or donut charts, particularly when there are more than two responses (i.e. rank ordinal responses rather than "will happen" / "won't happen"). The individual question-level will be displayed with a Y/N line chart over time.

Beyond displaying the WOC forecast in the aggregate, we are interested in evaluating forecaster reasoning and predictions, based on information available to them. Typically, the wisdom of crowds refers to an emergent phenomenon, such that forecasters make their initial decision with no crowd information, but the crowd's aggregate prediction is more accurate than individual forecasts. However, if forecasters make predictions with the crowd's prediction visible, this may influence or bias a forecaster's initial prediction. Thus, it is important not only to test the type of visualization used to display the crowd's forecast, but to display this information at an

appropriate time, to reduce confirmation bias or undue influence of the crowd. A future study will examine when information is provided in the Sphinx tool and the impact on forecast accuracy of the crowd, as well as on the justifications provided by forecasters for why they made their prediction.

7.0 REFERENCES

Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., & Van Gael, J. (2012, June). Crowd IQ: aggregating opinions to boost performance. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 535-542).

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. Computational Brain & Behavior, 1(1), 90-99.

Bobby, W. The Limits of Prediction—or, How I Learned to Stop Worrying About Black Swans and Love Analysis. *Studies in Intelligence*, *63*(4).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1), 1-3.

Forlines, C., Miller, S., Prakash, S., & Irvine, J. (2012, October). Heuristics for improving forecast aggregation. In 2012 AAAI Fall Symposium Series.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. Journal for Research in mathematics Education, 124-158.

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. Decision, 5(1), 1.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison.

Medical Decision Making, 31(3), 444-457.

Galton, F. (1907). Vox Populi. Nature, 75(1949), 450-451.

Haran, U., & Moore, D. A. (2014). A better way to forecast. California Management Review, 57(1), 5-15.

Iyer, R., & Graham, J. (2012). Leveraging the wisdom of crowds in a data-rich utopia.

Psychological Inquiry, 23(3), 271-273.

Kemp, M., & Kissane, B. (2010). A five-step framework for interpreting tables and graphs in their contexts. In: 8th International Conference on Teaching Statistics, 11 - 16 July 2010, Ljubljana, Slovenia.

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. American Educational Research Journal, 40(1), 281-310.

Landwehr, J. M., & Watkins, A. E. (1987). Exploring data. Dale Seymour Publications.

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

Lyon, A., & Pacuit, E. (2013). The wisdom of crowds: Methods of human judgement aggregation. In Handbook of human computation (pp. 599-614). Springer, New York, NY.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (214). The wisdom of select crowds. Journal of

personality and social psychology, 107(2), 276.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Murray, T. (2014).Psychological strategies for winning a geopolitical forecasting tournament. Psychological science, 25(5), 1106-1115.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. Journal of Behavioral Decision Making, 25(4), 390-401.

Peebles, D., & Ali, N. (2015). Expert interpretation of bar and line graphs: The role of graphicacy in reducing the effect of graph format. Frontiers in psychology, 6, 1673.

Ramakrishnan, N., Summers, K., Getoor, L., Srinivasan, A., Choudhury, T., Gupta, D., ... & Vullikanti, A. (2015). Model-based forecasting of significant societal events. IEEE Intelligent Systems, (5), 86-90.

Rufibach, K. (2010). Use of Brier score to assess binary predictions. Journal of clinical epidemiology, 63(8), 938-939.

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top- down and bottom-up processes. Topics in cognitive science, 3(3), 560-578.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. Journal of educational psychology, 91(4), 690.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception.

Journal of the American Statistical Association, 82(398), 454-465.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., & Baron, J. (2012, October). The good judgment project: A large scale test of different methods of combining expert predictions. In *2012 AAAI Fall Symposium Series*.

Walker, A. C., Stange, M., Dixon, M. J., Koehler, D. J., & Fugelsang, J. A. (2019). Graphical depiction of statistical information improves gambling-related judgments. Journal of gambling studies, 35(3), 945-968.

Wintle, B., Mascaro, S., Fidler, F., McBride, M., Burgman, M., Flander, L.,In & Manning, B. (2012). The intelligence game: Assessing Delphi groups and structured question formats.

8.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

ACE	Aggregative Contingent Estimation
CIA	Central Intelligence Agency
GJO	Good Judgement Open
HUMINT	Human Intelligence
IARPA	Intelligence Advanced Research Projects Activity
ISR	Intelligence, Surveillance, and Reconnaissance
PIR	Priority Information Requirements
SIGINT	Signals Intelligence
WOC	Wisdom of the Crowd