UNIVERSITY OF CALIFORNIA SAN DIEGO

An Analysis Of Judgment Variability Amongst Cybersecurity Participants When Asked To Forecast Cybersecurity-related Events

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Stephan Chenette

Committee in charge:

Professor Stefan Savage, Chair
Professor Geoffrey M. Voelker, Co-Chair
Professor Deian Stefan

2021

The Thesis of Stephan Chenette is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to acknowledge Stefan Savage for his guidance and feedback throughout the process of creating this thesis and for his support as my adviser and as the chair of my committee. Additionally, I would like to thank Geoff Voelker as my co-chair and Professor Deian Stefan who was gracious enough to join my committee in charge.

I would also like to acknowledge Ryan McGeehan for inspiring me to begin exploring the field of forecasting and eliciting expert opinion, and Patrick Shami for his guidance and expertise in the areas where I lacked a statistics background in order to make this thesis grounded and relevant.

Lastly, although not directly involved, I would like to thank Philip E. Tetlock for his writing on the subject of judgment forecasting. In particular, Mr. Tetlock's book, "Super Forecasting", and his research project, "Good Judgements Project", which further exposed and brought to my attention the lack of measurements within forecasting and offered trainable methodologies that could improve forecasting within an individual.

ABSTRACT OF THE THESIS

An Analysis Of Judgment Variability Amongst Cybersecurity Participants When Asked To Forecast Cybersecurity-related Events

by

Stephan Chenette

Master of Science in Computer Science

University of California San Diego, 2021

Professor Stefan Savage, Chair
Professor Geoffrey M. Voelker, Co-Chair

Forecasting using judgment is common in practice. From project planning to investing in a house, we routinely make important decisions based on how we expect the future to unfold. In much the same way, governments and businesses rely on forecasts related to political and economic events in order to appropriately plan for future challenges and opportunities. The cybersecurity industry is no different, and each year many security organizations release "annual predictions" of anticipated cybersecurity issues. The premise behind such predictions is that experts are in a superior position due to their knowledge and experience to anticipate future

cybersecurity-related events. However, this premise is untested and motivates this research.

This master's thesis examines the primary question of whether security professionals' predictions about future cybersecurity-related questions are systematically distinct from those of other Information Technology professionals, who lack the same specialized experience. In particular, this research examines the results of 20 security-related surveys. Each survey takes a different approach to the category and format of the question in order to analyze a variety of types of forecasting questions. Using a combination of measurement techniques, I determine if there exist any significant patterns among security professionals and non-security professionals. Additionally, I analyze two cohorts within security professionals to determine if there are measurable differences between them with regard to judgment forecasts. Through this study, I concluded that I cannot support the claim that security professionals offer a distinct judgment over non-security experts.

Chapter 1

Introduction

Cybersecurity has become a key topic for boards of directors and executive leadership today. In fact, cybersecurity-related risk is rated as the second-highest source of risk for an enterprise according to Gartner's 2020 Board of Directors Survey [3]. As a result, enterprises are forced to make key decisions about cybersecurity in order to minimize risk – both strategic (what to invest in, how much to invest, how to prioritize investment, and how to pivot as necessary to either anticipate significant changes in threat landscape or to react to a major event, etc.) and tactical – (regulatory considerations, what controls to adhere to and where controls can be placed to best mitigate risk to key assets, setup of configurations and policies, efficient and effective triage processes, etc). Today, it is unequivocally an art to do this well, and absent a clear science of cybersecurity decision making, we have tended to defer to experts. One of the underlying presumptions is that experts, because of their training and experience, have special insights into anticipating security issues and are thus able to make distinct and better predictions than non-experts. In this thesis, I focus on understanding the extent to which expert security predictions are indeed distinct from the predictions made by those less versed in the field.

For most of the cybersecurity industry, the input and the advice that cybersecurity leadership receives originate from similar sources. Some examples include the attendance of industry

events and presentations, vendors, consultants and consulting firms, analyst services, peer and industry groups, and even reliance on their own teams. Cybersecurity leaders, much like any executive leader, are expected to align their security program with that of their organizational culture, mission, business objectives, and regulatory compliance. Unlike many technology leaders, though, security leaders need to not only create a strategy and plan for their security program, but be able to quickly readjust their strategy and respond to unexpected potential threats in order to be able to answer to their board of directors, management, investors, and customers and to set expectations accordingly.

In the last 20 years, cybersecurity leaders have had to face new threats like distributed denial of service (DDOS), exploit attacks, ransomware, compromised websites, and supply chain attacks that have forced or led them to readjust priorities. In order to make appropriate decisions, they rely on experts to influence their decisions to an even greater extent.

More recently in 2020, all businesses had to adjust for the COVID-19 pandemic, and cybersecurity leaders had to alter their security practices to support an increasingly remote workforce and distributed infrastructure. This meant making key decisions around new software, new policies, and new or expanding infrastructure in order to support the business. This led to decisions around the protection of key business assets that previously were not as exposed to remote access but now had to be more accessible to support a remote workforce. In order to make appropriate decisions, cybersecurity leadership relied on experts to influence their decisions. The consequences of making the wrong decisions have led to regulatory fines, lawsuits, executives or employees being fired, loss of company credibility and reputation, company public stock dropping, and in some rare instances, companies going out of business [18] [9] [22]. These pressures and negative outcomes have made well-informed security decisions and practices even more crucial for company management and cybersecurity leadership.

Cybersecurity leaders today are typically provided with external advice from vendors or consultancies, who have specific commercial interests and who offer advice that is packaged with

marketing and advertising in order to communicate to organizations and cybersecurity executive leadership. This leads to flashy "and/or" exaggerations on the true state of the industry in order to stand out as a "thought-leader" in a crowded competitive landscape and vertical market. Each year, thousands of polls, whitepapers, and webinars are published from industry experts claiming a particular point of view on the state of threats, to which they conveniently offer the solution. This leads to conflict and obvious bias in the advice they offer, which makes it difficult for cybersecurity leadership to receive truly objective recommendations. Yet, cybersecurity leaders still accept this advice as it is what their peers follow, further perpetuating the cycle.

In 2020, the RSA Conference, the largest cybersecurity conference in the industry, had 658 exhibitors, which indicates how "noisy" the cybersecurity industry has become. In fact, this volume makes it hard to figure out where to find good advice. Some vendors will mix trends and statistics in their advice as a mechanism to subconsciously tie together both facts and subjective points of view as truth [19] [5]. In such reports, you will find statements mentioning supposed trends related to correlations of increased attacks on cloud infrastructure due to an increasingly remote workforce or to vulnerabilities in Internet of Things (IoT) devices due to the increased availability of 5G, along with semi-related statistics from previous years. These types of reports are indicative of the type of advice that is released by organizations and that cybersecurity leaders read, analyze, and make decisions from. Yet, rarely do those same leaders review advice, predictions, or claimed trends from previous years to determine how accurate or valuable such advice was. Accuracy is obviously one measure of the value of expert advice, but as was discussed by Robert Reeder, Iulia Ion, and Sunny Consolvo [17], a consensus is also an indication of value. Although accuracy and consensus are interesting points of study, this thesis addresses a different question.

In this thesis, I focus my work on the cybersecurity industry to review how forecasting is used in conjunction with expert opinion, and how different types of security individuals within the industry forecast based on certain types of questions. In particular, I will examine whether the

forecasting decisions made by security experts are distinct from those of non-security experts.

This research was inspired by the work from the original Good Judgement Project (GJP) research study [23] that showed "that the average expert had done little better than guessing on many of the political and economic questions" that were posed [24] and the more current ever-evolving Good Judgement Project [7] [25], whose goal is to determine whether some people are naturally better than others at prediction and whether prediction performance can be enhanced. Whereas the GJP explores the accuracy and methodology of forecasting, my research asks whether measuring security experts' predictions results in distinct forecasts or lack thereof, which would further emphasize the need within the security industry to provide in-depth training, analysis, accountability, and calibration. Recent articles demonstrate that the tech industry has some self-reflection on their own predictions and forecasts, such as the "Top Ten Worst Tech Predictions of All time" [21], but a true sense of accountability and reflexivity particular to cybersecurity predictions are lacking. This is insightful, particularly because it is commonplace for security companies to release predictions, and yet, they rarely revisit these predictions or make changes to hold themselves accountable.

This thesis takes relevant security forecasting questions of the kind used to drive decision-making and poses them to security experts to evaluate the similarity in their responses. Lack of consensus is an indicator of randomness and further corroborates the need for collaboration and accountability. The primary question this thesis sets to answer is if there exists distinction between security and non-security practitioners with regards to prediction forecasts.

In brief, after sending 20 different surveys that solicited decisions from security experts and non-security Information Technology professionals related to future cybersecurity questions, I found that security participants, who were of both strategic (Information Security leaders) and tactical (Information Security operators) backgrounds, and non-security Information Technology professionals had roughly no distinction in their forecasts. This was found to be true regardless of whether the questions were strategic or technical. In addition, I provided questions of different

types to determine if the format of the question had any relevance and found that independent of format, no distinction could be found between strategic security experts and technical experts or security experts and non-security professionals.

From the results of the research and surveys, this thesis' main findings and contributions are:

- A methodology for surveying security experts to forecast security-related questions. We review how and why our methodology for surveying changed over time and the methodology we used to compare three different cohorts: strategic security experts, tactical security experts, and non-security technical professionals. We also used a variety of question formats for participants to forecast. Some questions have traits of epistemic uncertainty, which is something an individual does not know but is, in theory, knowable. Whereas some questions have traits of aleatory uncertainty, which is something that not only does an individual not know, it is truly unknowable.
- An analysis of survey results across participants, and statistical methods to detect distinction amongst the different cohorts.
- A description of future work that could be researched and conducted based on this research's findings.

The rest of the thesis is organized as follows: Chapter 2 provides the necessary background for the basis of the thesis and reviews previous literature on the subject of expert security advice; Chapter 3 describes the methodology used in the analysis; Chapter 4 discusses the results and their implications; and Chapter 5 concludes the thesis and indicates future areas of research.

Chapter 2

Background

In this chapter, I discuss how decisions based on expert advice are used in the cybersecurity industry today and the risk and negative consequences of poor decisions made in the field. Additionally, I review some of the current literature on security advice by security experts as well as the use of forecasting and prediction for future events.

2.1 General Background

As discussed in the introduction, cybersecurity organizations and leaders rely on others' expertise and experience to make decisions and to prioritize projects and resources to effectively build and mature their security program with the ultimate goals of adhering to regulatory compliance and minimizing the risk of exposure for their organization. Expert advice primarily originates from external consulting firms, vendors, conferences, user groups, reports and literature, and internal personnel. The decisions that security leaders and operators make are in regards to both current and future circumstances and it is more common than not that an insufficient amount of information is available at the time of making decisions. Thus experts are relied on to make suggestions or provide foresight so that security leaders and operators can move forward with their decisions regardless of insufficient information. These decisions have significant business

impacts on an organization's share prices and revenue and may lead to legal actions when misled. Leaders and operators are rarely rewarded when security programs run without incident. However, if security programs fail or the security decisions prove to be incorrect, the outcomes in critical circumstances can lead to lawsuits, firings, professional reputation loss for the individual leader. Some examples of security decisions that can lead to these disastrous (and conversely positive) results include: which technologies, processes, and tools to invest in to minimize risk exposure, when and how often to patch and disrupt business processes, and how and when to balance security versus business usability. Security budgets are limited and resources are finite, so decisions and priorities must be made and set to meet those financial limitations. Ultimately, the role of any security leader is to minimize the risk of the business in order to accelerate business opportunities, yet an unfortunate and unintended consequence of minimizing risk can lead to a slow down in business. This is the balancing act security teams must constantly manage every day and why security leaders are under such high pressure to constantly tune their security program or keep up with the speed of the business while maintaining a level of acceptable risk. As cybersecurity has become increasingly a part of everyday business and cybersecurity incidents have become widespread news, the public has become more aware of the need for well-managed cybersecurity systems. In the past decade, multiple public organizational incidents and breaches have happened because of poor or incorrectly prioritized decisions. In hindsight, many of these incidents occur due to:

- Theft or loss of computers, laptops, portable electronic devices, electronic media, or paper files
- Insecure storage or transmission of Personal Identity Information (PII) and other sensitive information
- Hacked or revealed passwords
- Missing "patches" and updates

- Computers infected with a virus or other malware
- Insecure disposal and reuse of devices and/or materials
- Contractor computer compromised
- Development server compromised
- Application/infrastructure vulnerabilities and/or misconfiguration

In 2006, the U.S. Department of Veterans Affairs (VA) disclosed that a PC laptop and external hard disk containing personal data on 26.5 million veterans and active-duty military personnel were stolen from the home of a VA employee [4]. This was an eye-opening data breach that led to policy changes and regulations throughout the government from data encryption standards to data breach notification guidelines and data retention and minimization policies. Following this data breach, the Chief Information and Security Officers of government organizations were suddenly able to mandate changes, instead of simply making recommendations. Although many breaches have occurred in the 21st century, one of the most notable occurred in 2015 when the United States Office of Personnel Management (OPM), the agency that manages the government's civilian workforce, discovered that some of its personnel files had been hacked. Among the sensitive data that was exfiltrated were millions of records containing extremely personal information that had been gathered in background checks for people seeking government security clearances, including millions of records of people's fingerprints. The OPM breach led to a Congressional investigation and the resignation of top OPM executives. Its full implications are still being seen today for national security and for the privacy of those whose records were stolen. Data breaches occur in all industries and sectors. In the private sector, one very well-publicized example is the 2013 Target Breach. At the time, Target had a very experienced security leader and security team and had invested enormously in controls, which had logged the anomalous adversarial behavior. Despite these controls, the incident response security team was not alerted

in time before irreparable damage was done. Attackers were then able to utilize stolen third party credentials to exploit weaknesses in Target's system, access a customer service database, install malware on the system, and capture full names, phone numbers, email addresses, payment card numbers, credit card verification codes, and other sensitive data because of the multiple policy and technology decisions the security leadership and team had made. The consequences of the Target Breach resulted in Target paying $18.5 million to 47 states and the District of Columbia as part of a settlement that determined that the breach had compromised the data of millions of customers [14]. In 2017, the Russian military launched a ransomware attack known as NotPetya. The planted ransomware paralyzed multinational companies and permanently locked tens of thousands of computers internationally. Even in 2021, it is still considered the most destructive and costly cyberattack in history. For over ten years, ransomware and its capabilities as a particular attack vector have been a subject of great debate amongst organizations and security teams as future decisions must be made to combat such threats. Yet ransomware is only one vector that causes significant damage to businesses. In 2020, Russian state threat actors successfully compromised the U.S. security company, Solar Winds, and implanted a backdoor in their update software, giving the Russian hackers access to monitor and remotely penetrate public and private companies. This incident demonstrates the multiplicitous ways attackers use 3rd party access to gain direct control of critical systems. Although the U.S. government placed sanctions against such state actors, organizations of all sizes were forced to discuss internally the future decisions that must be made to minimize future business risk with expert team members, consultants, and trusted parties.

Today, occurrences of data breaches have become regular topics for headlines in the media. It is one reason why the World Economic Forum (WEF) ranks cybersecurity amongst the top ten risks of immediate concern [26]. Despite security leaders and teams' dependence on expert opinions to plan and make decisions in order to best be prepared for future threats, organizations continue to be breached. This troubling fact suggests that the advice experts offer

does not adequately address cybersecurity's evolving needs or that experts lack the appropriate credentials and expertise or are more unreliable than they seem. Part of the motivation of this research is to question whether organizations should rely on expert opinion for future decisions and how much value they should place in these opinions.

2.2 Related Research in Expert Advice

Previous literature on expert opinions and forecasting using their judgement exists both in regards to cybersecurity and non-cybersecurity topics [19] [5] [12] [8] [16] [2] [1]. In general, forecasting using judgment is a common practice. Every day, we, as individuals, businesses, and corporations, consume expert opinions and judgments, which we use to make our own subsequent decisions. Judgments using forecasting have been applied to many situations and multiple industries as well. One well-known example in which expert judgment was applied to political decision-making was during the lead-up to the 2003 Iraq Invasion by the U.S. Military. In this case, the U.S. intelligence community was asked by the U.S. government to determine the probability of weapons of mass destruction (WMD) in Iraq. Based in part on their answer, which incorrectly implied Iraq possessed WMD, the U.S. invaded Iraq and overthrew Sadaam Hussain.

Following the Iraq Invasion, the intelligence community took a particular step forward in the field of forecasting and created the Intelligence Advanced Research Projects Activity (IARPA) in 2006. Its mission was and is to fund cutting-edge research with the potential to make the intelligence community smarter and more effective, particularly in its forecasting of future events. Then, in 2008, the Office of the Director of National Intelligence, which sits atop the entire network of sixteen intelligence agencies, asked the National Research Council to form a committee. The task was to synthesize research on good judgment and help the Intelligence Community put that research to good use.

In 2010, the Office of Incisive Analysis (OIA) at the Intelligence Advanced Research

Projects Activity (IARPA) created the Aggregative Contingent Estimation (ACE) program and tournament, which ran from June 2010 until June 2015 [11]. The Goal of the IARPA-ACE tournament was "to dramatically enhance the accuracy, precision, and timeliness of intelligence forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts" [11]. The website claims that ACE seeks technical innovations in the following areas [11]:

- Efficient elicitation of probabilistic judgments, including conditional probabilities for contingent events
- Mathematical aggregation of judgments by many individuals, based on factors that may include: past performance, expertise, cognitive style, metaknowledge, and other attributes predictive of accuracy
- Effective representation of aggregated probabilistic forecasts and their distributions.

In 2011, The Good Judgment Project (GJP) was assembled as a team participant in the Aggregative Contingent Estimation (ACE) program in collaboration with IARPA-ACE. GJP was one of many entrants in the IARPA-ACE tournament and has repeatedly emerged as the winner in the tournament [10]. A commercial spin-off of GJP [7] started to operate on the web in July 2015, and today has several thousand participants. Participants are now able to post questions and make judgments on existing questions related to economic, political, and industry-specific questions every day and week. Examples of such questions are [7]:

- What will be the end-of-day price of Cardano's Ada cryptocurrency on 1 July 2021?
- What will happen next regarding the price of bitcoin?
- Will the powers of the government of Myanmar cease to be held by the military before 5 February 2022?

- How many people in the U.S. will have received one or more doses of a COVID-19 vaccine as of 28 February 2021, according to the CDC?

Although much progress has been made in the field of expert judgments and forecasting, the pioneering scientific research conducted by Philip E. Tetlock, who also helped lead the GJP, is still integral for understanding how expert forecasting and judgment is applied to cybersecurity 40 years later. Tetlock's research [23], which was conducted between 1984 and 2004, was one of the most comprehensive analyses of expert prediction ever conducted. The research study assembled a group of some 280 anonymous volunteers of economists, political scientists, intelligence analysts, and journalists whose work involved forecasting to some degree or other. These experts were then asked about a wide array of subjects, and in all, made some 28,000 predictions over that thirty-year period. Once the relevant date of the subject of the forecasting question passed, the veracity of the predictions was determined, the data analyzed, and the average expert's forecasts were revealed to be only slightly more accurate than random guessing.

One key conclusion from Tetlock's research was that a group of experts that clearly exhibited a particular style of thinking existed, compared to the body of experts as a whole. This particular group explained their forecasts in probabilities, was less anxious as to whether they had 100% confidence in a prediction, and were more comfortable discussing uncertainty [23]. It is important to acknowledge that this experiment involved individuals making subjective judgments in isolation, in contrast to GJP and ACE judgments, which allowed for forecasting amongst team consensus.

Almost 20 years after Tetlock's research study was concluded, his research initiatives have led to projects, such as GJP, which have noted that the top forecasters in GJP are "reportedly 30% better than intelligence officers with access to actual classified information" [20]. This is a true systematic improvement. What is noteworthy about projects like the GJP is that in using a team rather than an individual to forecast future outcomes, they improve their predictions dramatically. This body of research suggests that there are two important factors when forecasting: that an

individual's specific style of thinking changes the accuracy of predictions and that having a team work on predictions improves the outcomes compared to a sole person's predictions.

So what has changed in the last 40 years and what have we learned about the format and structure of forecasting questions? Why have forecasters definitively shown to forecast more accurately since Tetlock's initial study? Lawrence et al. [13] have shown that the accuracy of judgmental forecasting improves when the forecaster has: (i) necessary domain knowledge; and (ii) more timely, up-to-date information. In addition, GJP research [23] has shown that harnessing a blend of statistics, psychology, subject-specific training in a forecaster's professional domain, and discussion and collaboration between individual forecasters consistently produced the best forecast.

With this advancement in forecasting knowledge, techniques, and methodology, one would assume that most industries have trained their experts and are now able to predict more accurately or at least demonstrate some sense of consensus when forecasting. This is particularly important as so many key decisions are based on reports released by industry-leading organizations that provide intelligence in order to aid decision-making.

2.3 Related Research in Cybersecurity-Related Expert Advice

A review of previous literature to explore areas of research specifically within cybersecurity expertise and advice has shown expert advice as less impactful than potentially assumed. Previous research findings have shown that such advice is sometimes not followed, unconvincing, and/or overwhelming in some cases, inaccessible. Previous literature has also shown acceptable accuracy related to predicting risk but relies on previous data being utilized [12] [8] [16] [2] [1].

A study by Ion et al. in 2015 [12] exposed the disconnect between the advice from security experts and non-security-experts and how experts might follow their own advice, but non-experts do not follow their own advice. Another study from C. Herley, published in 2009 [8], found that

most security advice simply offers a poor cost-benefit tradeoff to users and is rejected because it is a daily burden, which must be applied to the whole population, whereas the harm suffered is only by the small fraction that becomes victims annually. Herley's research explains several factors as to why users reject security advice: one being they are overwhelmed by the amount of advice being given, and the second being they are not convinced that a particular action or decision has an equal or greater tradeoff economically. In 2016, Redmiles et al. [16] found that there exists a socio-economic "digital divide" and unfair difference as to which advice sources users have access to, leading to a difference in security behavior and beliefs when encountering potentially malicious cybersecurity behavior. Additionally, this research found that certain users put more trust in the source of the advice rather than the content of the advice, rejecting equally or more valid advice from other expert sources that were less known or prestigious.

In 2014, Canali et al. [2] demonstrated the effectiveness of risk prediction and provided results that showed that it is possible to predict with reasonable accuracy the users that are more likely to be the victims of web attacks by analyzing their browsing history. Similarly, Bilge et al [1] found by analyzing binary file appearance in logs of machines, the risk of infected machines can be predicted and calculated months in advance. An important factor here is that previous data was critical in effective and accurate risk prediction. Yet neither study addressed predictions with limited data or information and the predictions' accuracy.

Although previous literature exists on expert advice and its applicability, how the advice is interpreted, and experts' capabilities for predicting future risk, little effort, however, has gone toward examining whether security professionals' predictions about future cybersecurity-related questions are systematically distinct from those of other Information Technology professionals when providing such advice. As the body of literature shows, measuring distinction amongst experts and non-experts is not well studied and is a needed area of research. Thus this research study attempts to remediate that gap and determine what distinction, if any, exists among experts and non-experts when forecasting.

Chapter 3

Methodology

As addressed in previous chapters, research has been done on the accuracy of experts' forecasts, but little research has been done on what distinction exists among unique groups of experts and non-experts. My research aims to address that lack through the work presented here. To better examine if patterns of distinction amongst security experts and non-security experts exist when forecasting security-related questions for future events, I set up a research study that consisted of a series of forecasting questions that were sent to participants over a period of 4 months and analyzed the results as follows. In this chapter, I will explain the methodology by which this research was conducted.

3.1 Data Collection

The research study consisted of two parts: a pilot study, designed to assess the question and presentation format, and the primary study, which then used a fixed survey format for all questions. In the pilot study, a small set of security participants volunteered to participate. Each participant was told that this research study was for a master's thesis and that the thesis focused on soliciting data to determine patterns amongst security and non-security individuals regarding forecasting future security-related events. Surveys 1-6, which were the first group of surveys

compiled for the pilot, were sent to an estimated 25 participants that I had personal relationships with. The participants provided feedback through email on the question format, which allowed for the improvement and finalization of the question format for the primary study.

For the primary study, the number of security and non-security participants was expanded and a newly revised survey format was sent to an estimated 180 participants. The participants in the primary study were either direct connections from industry events or working security groups or they were introduced by direct connections. Some of these participants were of a security background and some were of a non-security background. Of those that were of a security background, I intentionally solicited strategic leaders as well as tactical operational practitioners for participation.

Direct security connections that were invited were individuals who demonstrated expertise in the field of cybersecurity and who had first-hand knowledge of their particular cybersecurity and professional backgrounds. Examples of individuals with such backgrounds that participated in the survey included Chief Information Security Officers (CISOs) from Fortune 500 companies, I.T. Security Senior Engineers for top-tier consulting firms and organizations, and vulnerability and exploit researchers. In general, though, because I could not and did not personally confirm the backgrounds of all our participants, I added a series of demographic questions that were appended to the survey with the intention of placing the participant into a cohort of strategic leadership vs operational practitioner, as well as if they associated themselves with security or non-security operations for their day-to-day job. It is important to note that each participant self-selected their group and security background and the validity of their self-selection was not investigated. At its peak, the research study solicited approximately 180 participants for Surveys 10-20, but the official research study analysis considered Surveys 7-20. At no point in the research study was any participant paid or promised anything in return for their participation.

Within the research study, each survey had one primary question each participant was asked to forecast and answer for. The intention with the question format was to select subjects

that security strategic leaders and operational practitioners would likely encounter in their day-to-day jobs. The surveys also considered that when these types of questions come up, relevant information would likely be paired with such questions. In the preliminary part of the study, I initially provided sources of information for participants to use as reference and/or offered multiple choices. Following the feedback received, I changed the question format and eventually removed any relevant information as well as eliminating multiple choice options in the official research study to reflect a realistic scenario where these types of questions come up, as individuals are not typically provided multiple choice answers or sources of past data when being posed a question. I also opted to remove the requirement that the questions posed to participants needed to have a verifiable answer. In reality, many cybersecurity questions that are presented to strategic leaders and tactical operational practitioners are not always verifiable amongst the industry as the data is private and not shared.

The overall nature of the survey questions fluctuated each week from strategic to operational. Strategic questions related to questions that at a top-level, security executives would encounter and had an overall likelihood to be discussed at a board of directors, executive, or management meeting. Operational questions in comparison related to questions different types of tactical security practitioners would encounter at a technical level that related to priorities of patching vulnerable machines or relevant threats.

The eventual finalized question and presentation format solicited the following information from each participant:

- An answer to the question
- Their individual forecasting approach
- An answer to the question as a lower and higher range with a 70
- Level of research time and effort in providing an answer

- Level of intuition used vs research time in providing an answer
- Background information (which was later used to divide participants into strategic or operational and security and non-security)

As an example of the eventual format of a survey, here is the layout of Survey 11:

Table 3.1: Layout of Survey 11

Question	How many years until the general public learns of an instance of a processor side-channel attack (e.g., Spectre/Meltdown, Zombieload) that was leveraged by a blackhat attacker against cloud infrastructure (i.e. AWS, GCP, Microsoft Azure) of a public company to perform a VM escape or attack shared virtualization resources? Note: (Forecast an exact number, e.g. x, where x is a whole number)(Forecast a range e.g. y-z, where you have 70% confidence that the correct answer will fall within your forecasted range. The format of your answer should be two whole numbers signifying a range e.g. y-z)
Forecast an exact number	[Participant quantitative answer]
Forecast a range	[Participant quantitative answer in the format y-z]
Can you share details as to your approach and methodology for how you came up with your forecast?	[Participant qualitative answer]
Describe your Forecasting Methodology (from 1-5, where 1 is a little and 5 is a lot)	[Participant quantitative answer from 1-5]
How much of your forecast was based on intuition/experience (from 1-5, where 1 is a little and 5 is a lot)?	[Participant quantitative answer from 1-5]
How much of your forecast was based on having to do research for this particular question (from 1-5, where 1 is a little and 5 is a lot)?	[Participant quantitative answer from 1-5]
If you did perform research, how much time did you spend on research (from 1-5, where 1 is a little and 5 is a lot)?	[Participant quantitative answer from 1-5]
How would you characterize your professional background? (Sliding scale to allow you to characterize your role as a scale) (left to right) left most is Leadership/Strategic, right-most is Practitioner/Operational/Tactical	[Participant quantitative answer from 0-5]
Do you have a background in Information Security (AKA Cybersecurity)? (from 0-5, where 5 indicates you have a professional or equivalent background)	[Participant quantitative answer from 0-5]

Each survey was distributed directly via email BCC'ing each participant using my personal gmail.com email address. The body of the email included a reminder of what the research study was for, a brief explanation of the question, and a link to surveymonkey.com, which was used as the mechanism to collect answers and to track and organize the survey data. I later normalized the data for analysis. It is worth noting that I chose to directly email each participant because, in the preliminary feedback stage, there were continued issues with survey emails landing in spam folders of participants when using SurveyMonkey for distribution and resulted in a huge drop in the rate of participation for those initial surveys.

As evident in Figure 3.1 (below), I increased the number of participants at Surveys 9 and 10, and when I received a lower response rate, I re-sent the surveys as a reminder. The goal was to consistently send new surveys weekly or biweekly.

3.1.1 Survey Participation

Survey Participation

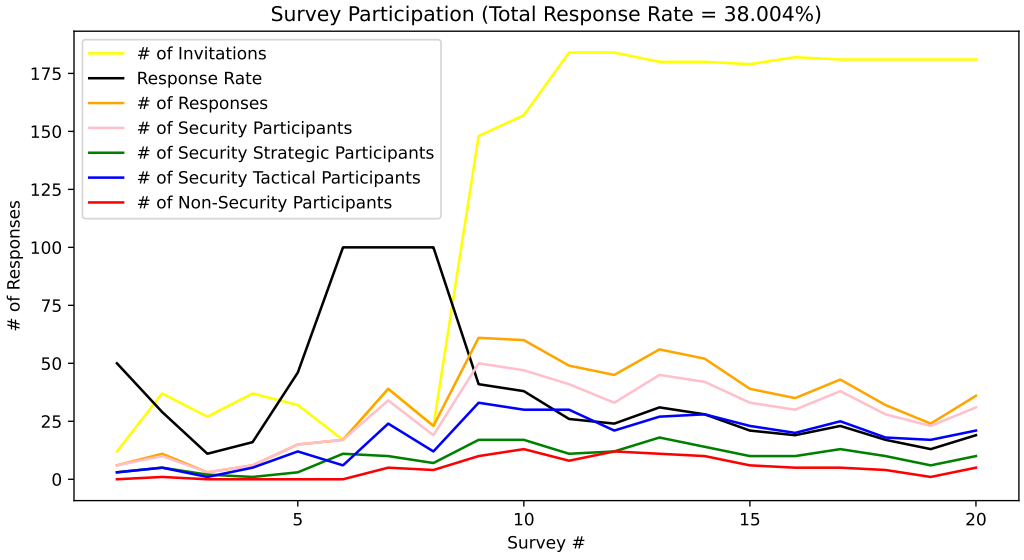

Figure 3.1: The Total Number of Survey Invitations vs The Total Number of Responses Over the Timeline of the Research Study

Because of the nature of the questions, I had speculated that security questions and technical questions would be intimidating for many participants and they likely would not answer. Based on anecdotal feedback from participants, that assumption was accurate. The overall response rate averaged 38% from Surveys 1-20. In general, there was a higher response rate for strategic questions over operational questions.

For the cadence of the surveys, emails were sent out at the start of the week, giving each participant an entire week to answer, and then the surveys were actively closed at the end of that week. Participants were not able to see the answers of other participants or change their answers once the survey was closed. In some cases, surveys were left open if there was a low rate of responses and answers. When there was a low response rate for a specific week's survey, all participants were sent a reminder email, which resulted in a small uptick in participation.

3.2 Analysis

The primary goal of the survey responses was to analyze and compare the survey participants against one another to determine if distinction existed with regard to forecasted judgments. The entire participant dataset was divided into the following cohorts:

- A security cohort (comprised of a separate security strategic leadership cohort and a separate security tactical operational practitioners cohorts)
- A non-security cohort (others who did not fall into the parameters of the security cohort)

Security strategic leaders were compared to security tactical operational practitioners and all security participants were compared to all non-security participants. We were able to divide participants into these cohorts because in each survey, participants were asked to describe their security and professional background. Participants scored themselves within a range of 1-5 to describe their day-to-day job as either a tactical operational practitioner or a strategic leader.

Those who marked "1" had the least relevance or involvement in those jobs or tasks, whereas those who marked "5" held positions that required the most expertise in those areas. We also asked our participants to score their security experience using a range of 1-5, with "1" being the least experienced and "5" the most, which was used to divide our participants into the security and non-security cohorts.

To determine if forecast answers between participants were distinct, we first ran normality tests for each survey and then ran the appropriate hypothesis tests depending on whether the dataset met the criteria for normal distribution. Normality tests were used on each survey to determine if each dataset was well-modeled by a normal distribution (Gaussian distribution). To determine normal distribution, several normality tests were considered, such as Anderson-Darling, Cramer-von Mises, Lilliefors, and Shapiro-Wilk, but ultimately Shapiro-Wilk was selected because it was the most appropriate for the dataset, as it did not require knowing the mean and variance a priori, it works well with smaller data sets, and previous research indicates that the Shapiro-Wilk test is the most powerful normality test [15] [27].

For datasets that failed to meet the criteria for normality ($p \leq 0.05$), non-parametric and parametric hypothesis tests were used to compare the datasets. In many cases, the dataset for a particular survey was smaller than what was required of particular non-parametric tests (e.g. $\geq$ 20 participants for Mann-Whitney). In these cases, we performed multiple hypothesis tests using different algorithms and looked for consistency among their results. In particular, we used the following non-parametric tests for non-normally distributed datasets:

- The Mann-Whitney U test
- The Kruskal-Wallis H test
- Mood's Median test

There were several other non-parametric tests that were considered and ruled out as inappropriate. For example, the Wilcoxon signed-rank test was considered, but it required

comparing paired data samples, and our dataset was not paired. The Friedman Test was also considered, but it required the comparison of more than two datasets and this study was only comparing two datasets at a time.

Of the nonparametric hypothesis tests, three tests were selected: Mann-Whitney U, Kruskal-Wallis H, and Mood's Median. The Mann-Whitney U test was selected because it supports non-normal datasets. It is important to note that the Mann-Whitney U test requires that each dataset have a sample size of 20. Because many of our surveys and the resulting cohort datasets did not meet this last criterion, I opted to include additional nonparametric hypothesis tests for comparison, namely the Kruskal-Wallis H test and Mood's Median test. I then used the results of all three tests to assess the similarity or difference between each pair of datasets.

An important decision made when running the Mann-Whitney U and Kruskal-Wallis H tests is that both tests were run as two-tailed hypothesis tests [6] rather than one-tailed tests. A one-tailed test looks for a directional relationship, i.e. whether the values in one dataset are larger than those in the other. A two-tailed test is appropriate when we simply wish to know whether two datasets are different, regardless of which contains larger values. This study seeks to determine whether the different cohorts produce significantly different forecasts, and does not seek to determine which cohort produces the larger forecast values. Therefore, the two-tailed versions of each test are the appropriate ones for this research. In addition, we had planned to use the Welch-Satterthwaite version of the Student's T-Test for those surveys which met the criteria for normality (according to the Shapiro-Wilk or Lilliefors tests). The student's T-test is a parametric hypothesis test, which assumes that the data are normally distributed. The Welch-Satterthwaite version of this test does not assume the variances of the two cohort datasets are equal. Since we do not have reason to assume equal variances, this test would be appropriate if we had found evidence of normality within any of the surveys. However, since the Shapiro-Wilk and Lilliefors tests found non-normality in all cases, we had no occasion to use this test.

Ultimately, for each survey, four datasets were computed in order to determine if a

significant difference existed between:

- The security strategic leadership cohort vs the security operational practitioners cohort
- Security cohort vs non-security cohort (the security cohort was comprised of the security strategic leadership and the security tactical operational practitioners cohort)

The following data from those cohort datasets were analyzed and compared for each survey:

- Forecast answer
- Forecast lower range answer
- Forecast upper range answer
- The numeric range between Forecast lower range and upper range answers
- Research time
- Research effort
- Intuition level

For the comparison of each dataset, normality and significance for each dataset within each survey was computed. In addition, visual staircase histograms were built to enable visual analysis of the datasets and to make conclusions. The results and discussion of the survey results can be found in “Chapter 4: Results”.

Chapter 4

Results

In the previous chapters, I discussed the driving hypothesis for this study, which was to determine if there is distinction amongst security and non-security experts when forecasting relevant security questions. Here, in Chapter 4, I present the surveys used and describe the findings from those surveys.

4.1 Participant Invitations and Survey Response Rate For the Surveys

The initial surveys (Surveys 1-7) were a preliminary measure so as to set the format of the questions appropriately. Once the format was finalized, the majority of participants from both security and non-security backgrounds were invited to complete the new surveys.

4.2 Participant Background

Surveys asked participants to self-describe two characteristics of their background:

- Strategic or Tactical (scored 1-5)

- Security or Non-Security (scored 1-5)

To determine which participants fell into the Strategic or Tactical categories, responses of 1 or 2 were grouped together and labeled “Strategic” and responses of 3-5 were grouped together and labeled “Tactical.” To determine who was in the Security or Non-Security groups, responses of 1-3 were grouped together and labeled “Non-Security” and responses of 4 or 5 were labeled “Security”.

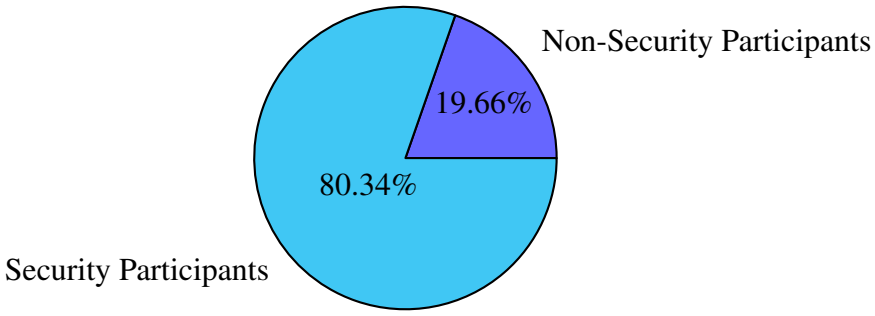

Figure 4.1: Security vs Non-Security Participation

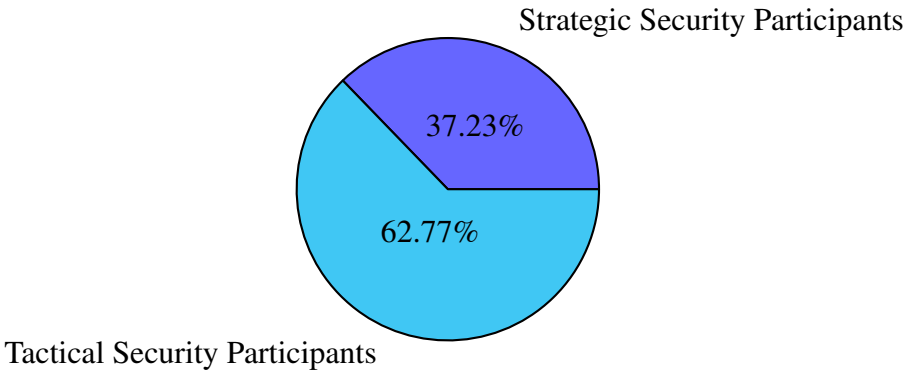

Figure 4.2: Strategic Security vs Tactical Security Participation

4.3 Breakdown of Type of Survey Security Questions

Within the surveys, both strategic and technical/tactical questions were posed to allow an analysis of whether the type of question had an impact on distinction between groups.

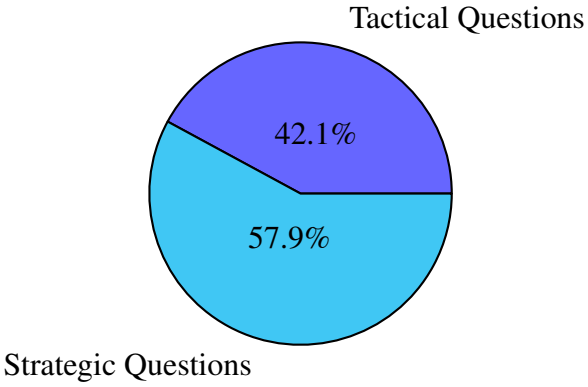

Figure 4.3: Percentage Breakdown of Type of Security Question Across All Surveys

4.4 Survey Question Format and Participant Summary

Strategic = Strategic/Leadership

Tactical = Tactical/Technical

R/I = Responses / Invitations

SEC = Security Participants

SECSTR = Security Strategic Participants

SECTAC = Security Tactical Participants

NONSEC = Non-Security Participants

Table 4.1: Survey Questions and Details

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#1 05/04/2020	How many companies will disclose a data breach to the California State Attorney General in 2020?	Strategic	6/12	6	3	3	0
#2 05/11/2020	How many unique CVEs will be in the June 2020 Google Android Operating System public security bulletin?	Tactical	11/37	10	5	5	1
#3 05/18/2020	What will the average enterprise ransomware payment be in Q3 of 2020?	Strategic	3/27	3	2	1	0
#4 05/18/2020	How many unique CVEs will be in the June 2020 Google Android Operating System public security bulletin? (Duplicate/Improved)	Tactical	6/37	6	1	5	0
#5 05/25/2020	In the next 12 months, which of the Fortune 100 companies will have a new public data breach mentioned in the New York Times first? (Note: Refer to the Fortune 100 list as of May 25th, 2020) (2nd Note: Digital or Print Edition)	Strategic	15/32	15	3	12	0

Table 4.1 Survey Questions and Details, Continued.

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#6 05/25/2020	For the future date of January 1st, 2021, what will be the largest number of individuals notified by a single company/entity documented in the Department of Health and Human Services (HHS) breach portal? Note: At the federal level, companies that are required to comply with the Health Insurance Portability and Accountability Act (HIPAA) must both notify individuals when covered data is lost and report the incident to the Department of Health and Human Services (HHS). That information is then made publicly available at HHS's breach portal.	Strategic	17/17	17	11	6	0
#7 06/01/2020	By the end of 2020, How many Cybersecurity Series A investment deals will there be?	Strategic	39/39	34	10	24	5
#8 06/01/2020	By the end of 2020, What will be the largest public single payout by a Bug Bounty Program?	Tactical	23/23	19	7	12	4

Table 4.1 Survey Questions and Details, Continued.

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#9 06/08/2020	How many in the wild zero-day exploits for a Microsoft product will be publicly disclosed in Q4 2020 (October, November, December)?	Tactical	61/148	50	17	33	10
#10 06/15/2020	How many total GDPR Fines / Penalties for data breaches in f500 companies will be issued in 2020?	Strategic	60/157	47	17	30	13
#11 06/22/2020	How many years until the general public learns of an instance of a processor side-channel attack (e.g., Spector/Meltdown, Zombieload, etc) that was leveraged by a blackhat attacker against cloud infrastructure (i.e. AWS, GCP, Microsoft Azure) of a public company to perform a VM escape or attack shared virtualization resources?	Tactical	49/184	41	11	30	8
#12 06/29/2020	What will be the average tenure of a Chief Information Security Officer (CISO) by the end of 2030? (please form your answer as a number and indicate months or years)	Strategic	45/184	33	12	21	12

Table 4.1 Survey Questions and Details, Continued.

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#13 07/06/2020	How many years until Ransomware authors will send targeted deep fakes to ransomware targets? Further Details: Recipients will see realistic videos of themselves in compromising situations and will likely pay the ransom demand in order to avoid the threat of the video being released into the public domain. (please form your answer as a number and indicate years)	Tactical	56/180	45	18	27	11
#14 07/13/2020	How many years until the United States (U.S.) passes a Federal Law similar to the California Consumer Privacy Act (CCPA) or the General Data Protection Regulation (GDPR)? Further Information: The law would not necessarily need to have the depth of protection of GDPR. Rather, this is just a question about when the federal government will take its first steps to address the issue of protecting an individual's data across all industries holistically.	Strategic	52/180	42	14	28	10

Table 4.1 Survey Questions and Details, Continued.

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#15 07/20/2020	How many months from now until we see the next CVE for a Microsoft Office 365 Vulnerability?	Tactical	39/179	33	10	23	6
#16 07/27/2020	How many Fortune 1000 organizations will suffer a data breach in 2021 and fire or lose their CEO, CIO, or CISO within 12 months of suffering or publicly disclosing the data breach?	Strategic	35/182	30	10	20	5
#17 08/03/2020	What is the percentage likelihood that we will read publicly about attackers exploiting infrastructure that will lead to a DDoS or DoS condition affecting the 2020 U.S. National Elections?	Strategic	43/181	38	13	25	5
#18 08/10/2020	How many years until a typical cybersecurity budget for an organization surpasses 20% or greater of the total IT budget?	Strategic	32/181	28	10	18	4

Table 4.1 Survey Questions and Details, Continued.

Survey # / Date	Question	Type	R/I	SEC	SECSTR	SECTAC	NONSEC
#19 08/17/2020	Recent vulnerabilities were announced AKA Achilles vulnerabilities, detailing how flaws in Qualcomm Snapdragon chips could be exploited to monitor location and audio and to steal images and videos. They could also be exploited to render devices useless. The chips are used in hundreds of millions of Android devices. What is your percentage confidence that there will be a public story of these vulnerabilities (AKA Achilles CVEs) being publicly exploited before fixes are incorporated into the Android OS or Android devices that use Snapdragon?	Tactical	24/181	23	6	17	1
#20 08/24/2020	What percentage of Fortune 1000 companies will have a documented response plan for pandemics as part of their business continuity and disaster recovery strategy within 12 months from now?	Strategic	36/181	31	10	21	5

4.5 Official Research Study Results

Surveys 7-20 were considered for analysis and to draw conclusions as to whether or not we found evidence of a distinction between security experts and a difference in the predictions between security and non-security participants. As we can observe in Figure 4.7 and is described in detail in Section 4.6, the forecasting result, which was the main datapoint of the surveys, demonstrated that participants from both a security and non-security background lacked distinctiveness and security experts as a group including both strategic and tactical security background lacked consensus in their predictions. In addition to forecast results, I collected intuition level, research effort, and research time from each survey question, which were secondary data points. Analysis of those results demonstrated that there was a mix of distinctiveness between all survey questions. Further detail is described in section 4.6 at the individual survey question level.

Table 4.2: Selected Test p-values for Survey Forecast Results

Test	Survey #													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
All Participants (Lilliefors p-value)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.010037	0.001	0.119628	0.046222
All Participants (Shapiro–Wilk p-value)	0.000048	0.000001	0.0	0.0	0.0	0.004057	0.0	0.0	0.0	0.0	0.000549	0.0	0.28	0.93
Security and Non-Security Participants (Mann-Whitney-U p-value)	0.98	0.05	0.06	0.21	0.67	0.08	0.92	0.16	0.87	0.89	0.48	0.55	0.28	0.93
Strategic and Tactical Security Participants (Mann-Whitney-U p-value)	0.62	0.12	0.98	0.71	0.34	0.35	0.72	0.22	0.4	1.0	0.53	0.12	0.53	0.08
Security and Non-Security Participants (Kruskal-Wallis p-value)	nan	0.04	0.06	nan	0.66	0.08	0.92	0.16	0.86	0.87	0.47	0.53	0.25	0.91
Strategic and Tactical Security Participants (Kruskal-Wallis p-value)	nan	0.11	0.97	nan	0.33	0.34	0.71	0.22	0.39	0.98	0.52	0.11	0.5	0.08
Security and Non-Security Participants (Mood's Median p-value)	nan	0.3	0.3	nan	0.75	0.37	0.87	0.14	0.52	0.94	0.87	1.0	1.0	1.0
Strategic and Tactical Security Participants (Mood's Median p-value)	nan	0.59	0.85	nan	0.54	0.92	0.95	0.57	0.18	1.0	1.0	0.24	0.73	0.3

Table 4.3: Selected Test p-values for Survey Intuition Level Results

Test	Survey #													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
All Participants (Lilliefors p-value)	NaN	NaN	NaN	NaN	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
All Participants (Shapiro–Wilk p-value)	NaN	NaN	NaN	NaN	11	12	13	14	15	16	17	18	19	20
Security and Non-Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.0	0.000004	0.0	0.0	0.0000042	0.000003	0.0	0.0000014	0.000008	0.0
Strategic and Tactical Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.85	0.73	0.12	0.37	0.45	0.74	0.33	0.94	0.36	0.82
Security and Non-Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	0.01	0.02	nan	0.01	nan	nan	nan	0.05	nan
Strategic and Tactical Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	0.71	0.11	nan	0.44	nan	nan	nan	0.34	nan
Security and Non-Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	0.08	NaN	nan	0.08	nan	nan	nan	NaN	nan
Strategic and Tactical Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	NaN	NaN	nan	0.79	nan	nan	nan	NaN	nan

Table 4.4: Selected Test p-values for Survey Research Effort Results

Test	Survey #													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
All Participants (Lilliefors p-value)	NaN	NaN	NaN	NaN	0.001	0.001	0.001	0.001	0.001	00.001	0.001	0.001	0.002225	0.001
All Participants (Shapiro–Wilk p-value)	NaN	NaN	NaN	NaN	0.0	0.000001	0.0	0.0	0.003708	0.000006	0.0	0.000267	0.003476	0.0
Security and Non-Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.98	0.04	0.75	0.45	0.18	0.67	0.5	0.75	0.1	0.37
Strategic and Tactical Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.19	0.81	0.95	0.14	0.43	0.76	0.68	0.73	0.02	0.07
Security and Non-Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	0.03	nan	nan	nan	nan	nan	nan	nan	nan
Strategic and Tactical Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	0.79	nan	nan	nan	nan	nan	nan	nan	nan
Security and Non-Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	0.01	nan	nan	nan	nan	nan	nan	nan	nan
Strategic and Tactical Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	0.92	nan	nan	nan	nan	nan	nan	nan	nan

Table 4.5: Selected Test p-values for Survey Research Time Results

Test	Survey #													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
All Participants (Lilliefors p-value)	NaN	NaN	NaN	NaN	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
All Participants (Shapiro–Wilk p-value)	NaN	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.000001	0.0000036	0.0	0.0	0.001712	0.0
Security and Non-Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.51	0.28	0.01	0.04	0.06	0.61	0.18	0.13	0.1	0.21
Strategic and Tactical Security Participants (Mann-Whitney-U p-value)	NaN	NaN	NaN	NaN	0.28	0.55	0.34	0.01	0.95	0.96	0.95	0.4	0.8	0.17
Security and Non-Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Strategic and Tactical Security Participants (Kruskal-Wallis p-value)	NaN	NaN	NaN	NaN	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Security and Non-Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Strategic and Tactical Security Participants (Mood's Median p-value)	NaN	NaN	NaN	NaN	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

4.6 Result Details

While Surveys 1-6 were used for final survey design, Survey 7 was the first survey from which data was used for the overall analysis and conclusions of this research.

4.6.1 Survey 7 Result Details

The forecasting question asked in Survey 7 was: "By the end of 2020, How many Cybersecurity Series A investment deals will there be?" This question was strategic in nature, with a presumption that executives who track new and emerging security technologies and investment progression would have distinct responses. It is often the case that a Chief Information Security Officer or strategic leader will integrate emerging technologies into their security program and track potential candidate technologies. As noted in Table 4.2 and Figure 4.4, there were 36 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0.00048 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .98, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .62, indicating that the forecasts of the two groups were not significantly different.

Therefore, I conclude that given this particular strategic question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

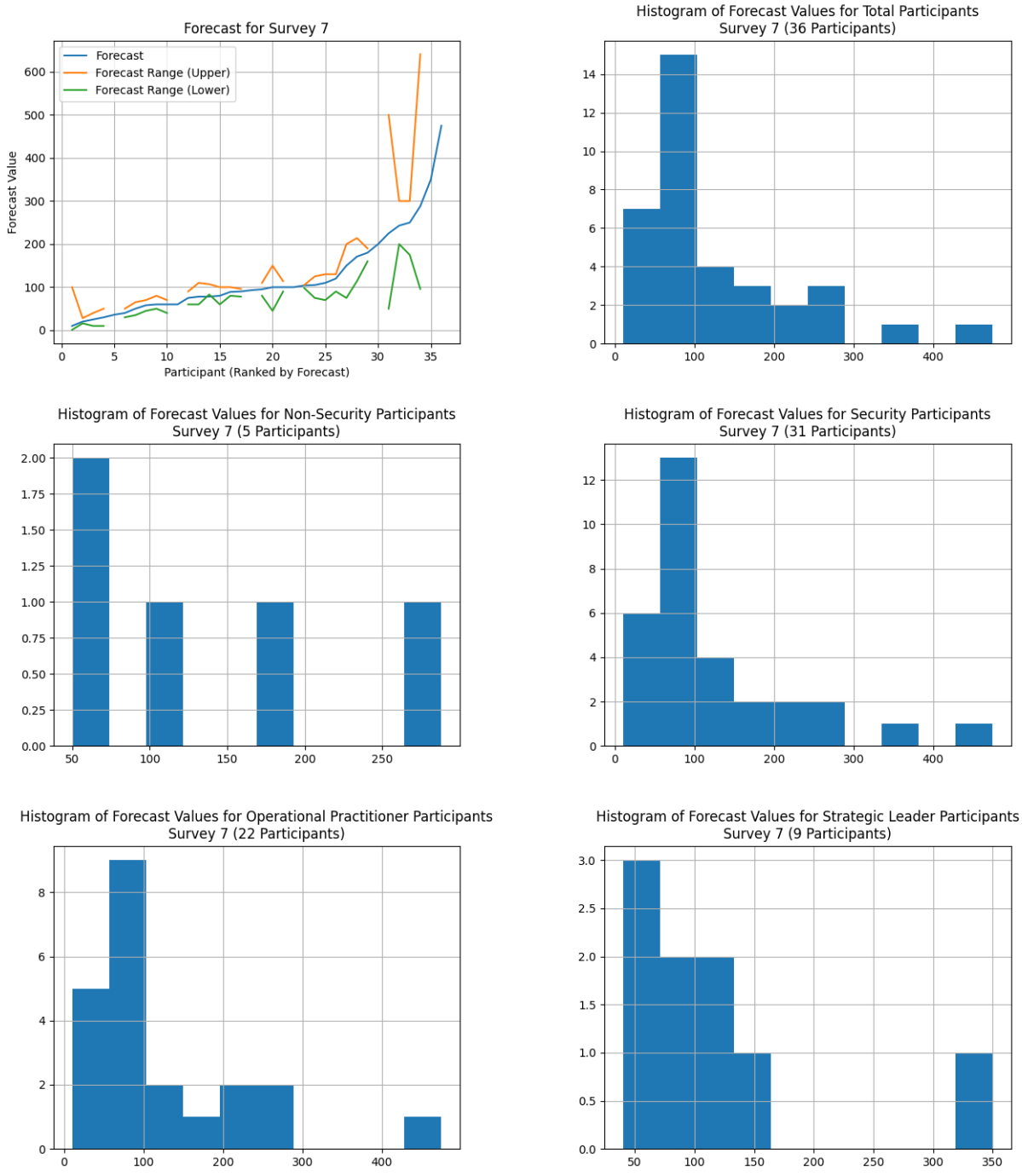

Figure 4.4: Forecast Results and Histograms of Forecast Results in Survey 7

Intuition Level, Research Effort, and Research Time Results

Survey 7 did not collect data on intuition level, research effort, or research time, and thus results do not exist to be analyzed. These data points were collected in Surveys 11-20 only.

4.6.2 Survey 8 Result Details

The forecasting question for Survey 8 was: "By the end of 2020, what will be the largest public single payout by a Bug Bounty Program?" This question was technical in nature with a presumption that technical security experts, who track bug bounty programs to determine vulnerability prioritization around attack surfaces, as well as focus their own efforts to find vulnerabilities and receive high paying bounties in return, would have distinct responses. As noted in Table 4.2 and Figure 4.5, there were 23 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0.000001, which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .05. A P-Value > 0.05 indicates that there is no evidence of difference in the predictions of the two groups, but in this case the P-Value is exactly .05. To get a second opinion, we performed the non-parametric Kruskal-Wallis test, and found a P-Value of 0.04, which is very close to the significance threshold of 0.05. Taking both of these P-Values into account, I conclude that the forecast results were on the cusp of distinctness, but not heavily indicative of a meaningful or significant conclusion. When using the Mann-Whitney test, the P-Value between the strategic security participants and tactical security participants was .12, indicating their forecast results

were not significantly different.

The results of this particular technical question were inconclusive with comparing the responses of non-security and security participants. However, a significant conclusion can be drawn from the responses of the strategic security participants and tactical security participants in that their forecasts were significantly similar, contradicting the typical belief that tactical security participants would be in a unique position to provide distinct forecasts related to this type of question over a strategic security participant.

Figure 4.5: Forecast Results and Histograms of Forecast Results in Survey 8

Intuition Level, Research Effort, and Research Time Results

Survey 8 did not collect data on intuition level, research effort, or research time, and thus results do not exist to be analyzed. These data points were collected in Surveys 11-20 only.

4.6.3 Survey 9 Result Details

Survey 9 asked: "How many in the wild zero-day exploits for a Microsoft product will be publicly disclosed in Q4 2020 (October, November, December)?" This question was technical in nature, with a presumption that technical security experts, who track published zero-day exploits to take action and communicate to management recommendations related to security resource and project prioritization, would have distinct responses. As noted in Table 4.2 and Figure 4.6, there were 61 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test and revealed that the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value when analyzing the security participants and non-security participants was .06, which is a P-Value > 0.05, which indicates that there is no evidence to show difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value when examining the responses from the strategic security participants and tactical security participants was .98, indicating that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular technical question, there was no significant distinction between the different types of participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

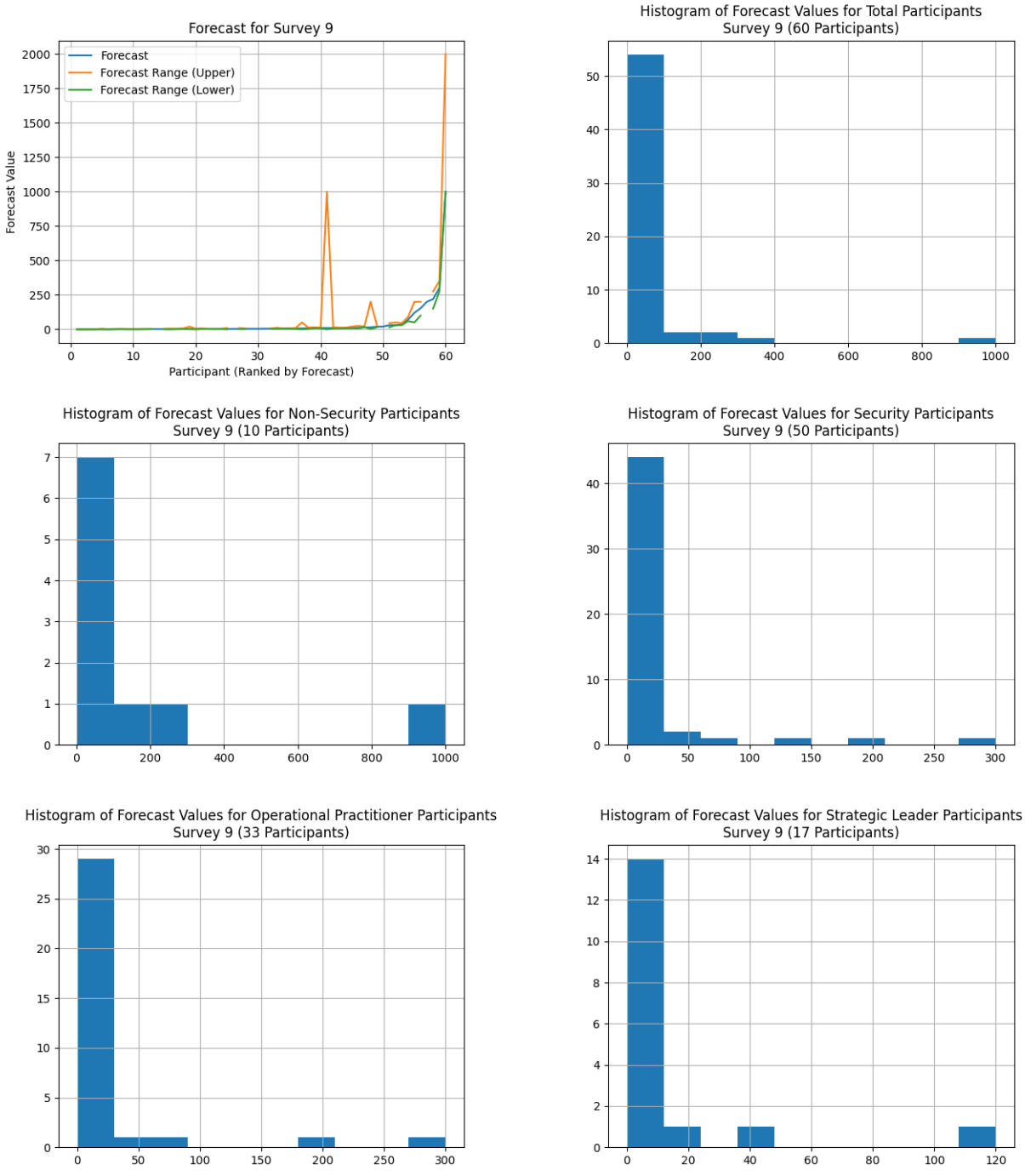

Figure 4.6: Forecast Results and Histograms of Forecast Results in Survey 9

Intuition Level, Research Effort, and Research Time Results

Survey 9 did not collect data on intuition level, research effort, or research time, and thus results do not exist to be analyzed. These data points were collected in Surveys 11-20 only.

4.6.4 Survey 10 Result Details

In Survey 10, forecasters were asked: "How many total GDPR Fines / Penalties for data breaches in f500 companies will be issued in 2020? Strategic in nature, this question presumed that strategic security experts, who have expertise regarding regulatory fines imposed on other organizations in order to prioritize their own resources effectively, would have distinct responses. As noted in Table 4.2 and Figure 4.7, there were 60 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .21, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the predictions of the strategic security participants and tactical security participants was .71, indicating that the forecasts of the two groups were not significantly different.

I therefore conclude that there was no significant distinction between participants for this particular strategic question, which is contrary to the assumption that a prediction given by a security participant would be distinctive from that given by a non-security participant.

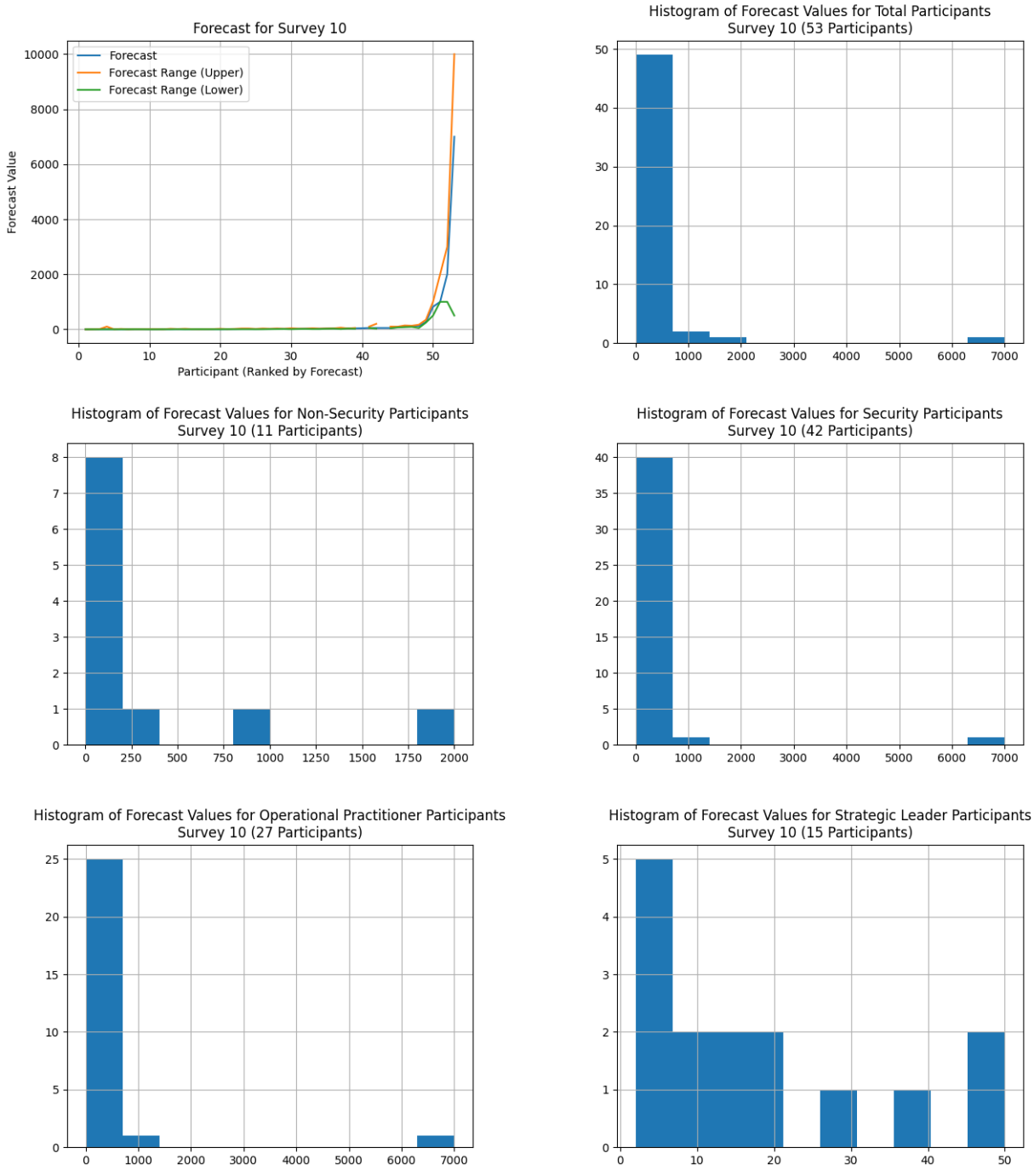

Figure 4.7: Forecast Results and Histograms of Forecast Results in Survey 10

Intuition Level, Research Effort, and Research Time Results

Survey 10 did not collect data on intuition level, research effort, or research time, and thus results do not exist to be analyzed. These data points were collected in Surveys 11-20 only.

4.6.5 Survey 11 Result Details

In Survey 11, participants were asked: "How many years until the general public learns of an instance of a processor side-channel attack (e.g., Spector/Meltdown, Zombieload, etc) that was leveraged by a blackhat attacker against cloud infrastructure (i.e. AWS, GCP, Microsoft Azure) of a public company to perform a VM escape or attack shared virtualization resources? This was a technical question in nature, with a presumption that technical security experts, who track and are involved with low-level security research related to vulnerability prioritization, would have distinct responses. As noted in Table 4.2 and Figure 4.8, there were 49 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Like the previous datasets that did not meet the criteria for normal distribution, I used the Mann-Whitney non-parametric test to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .67, which is a P-Value > 0.05 and indicates that there is no evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .34, indicating that the forecasts of the two groups were not significantly different.

In regards to this particular technical question, there was no significant distinction between

participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

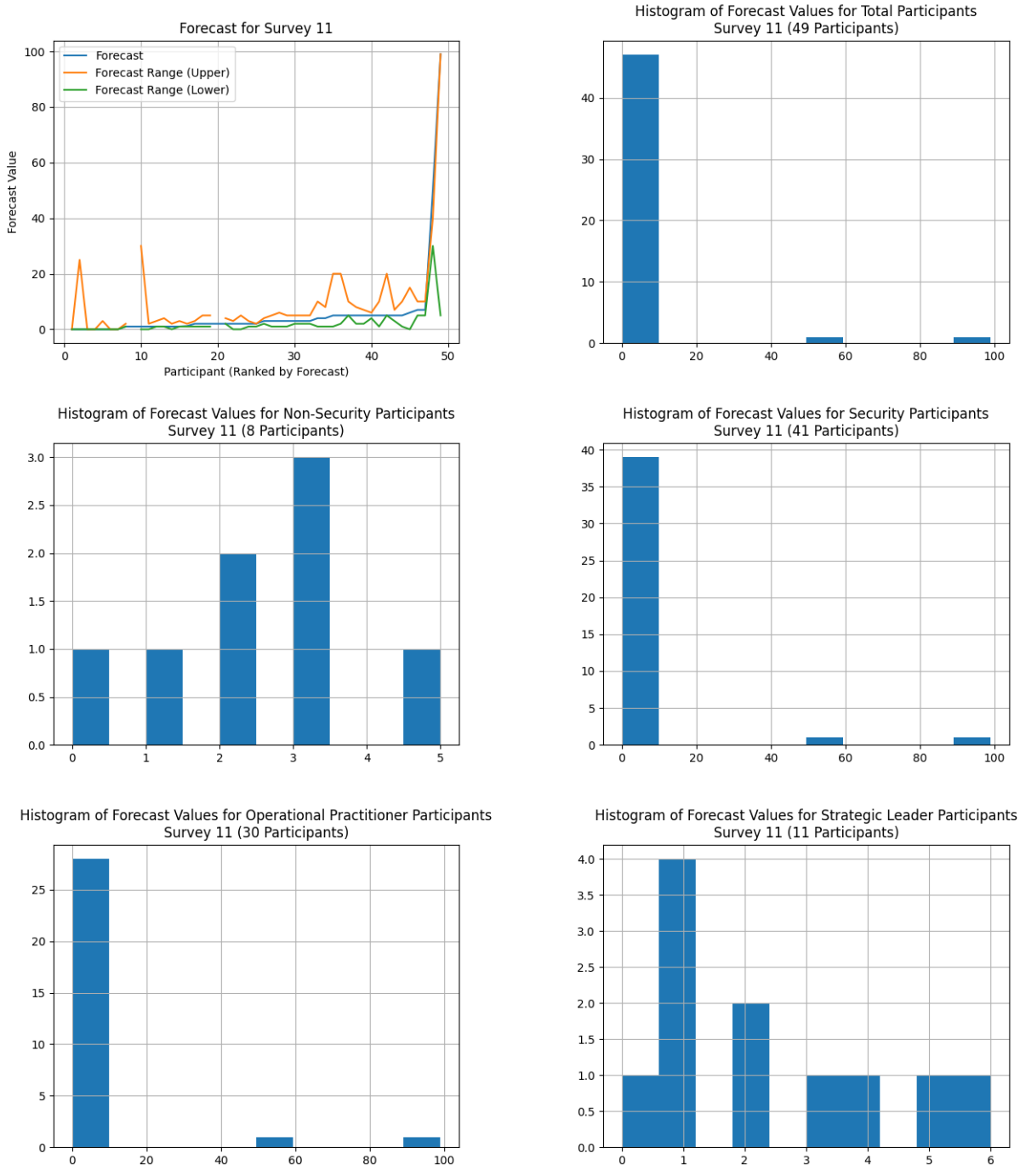

Figure 4.8: Forecast Results and Histograms of Forecast Results in Survey 11

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .74 for intuition level, 0.98 for research effort, and .51 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. For this particular question, the results demonstrated that there was no unique distinctiveness between non-security and security participants with regards to intuition level, research effort, or research time. The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .85 for intuition level, 0.19 for research effort, and .28 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. These results demonstrated that there was no unique distinctiveness between strategic security participants and tactical security participants with regards to intuition level, research effort, or research time.

4.6.6 Survey 12 Result Details

In Survey 12, participants were asked: "What will be the average tenure of a Chief Information Security Officer (CISO) by the end of 2030? (please form your answer as a number and indicate months or years)?" This was a strategic question in nature, with a presumption that strategic security experts, many who hold the job title of CISO or are on a career path towards such a role, would have distinct responses. As noted in Table 4.2 and Figure 4.9, 45 participants submitted responses to this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0.004057 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .08, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .35, demonstrating that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular strategic question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

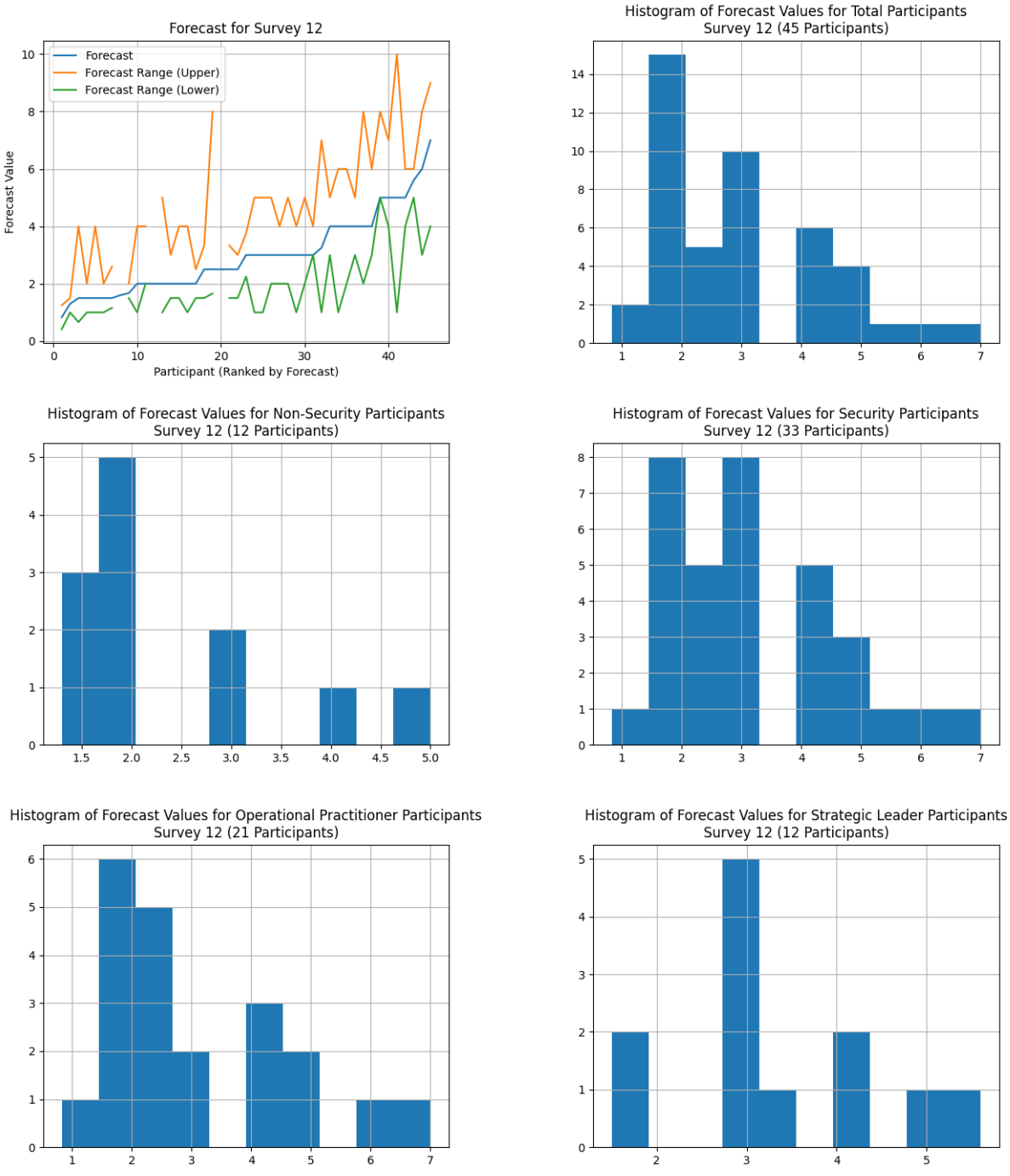

Figure 4.9: Forecast Results and Histograms of Forecast Results in Survey 12

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .01 for intuition level, 0.04 for research effort, and .28 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. The results demonstrated that there was a mix of distinctiveness for this question between non-security and security participants with regards to intuition level, research effort, or research time. Specifically, intuition level and research effort showed a distinction between non-security and security participants.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .73 for intuition level, 0.81 for research effort, and .55 for research time. A P-Value > 0.05 indicates that there is no difference in results between the two groups. So for this particular question's results, there was no unique distinctiveness between the responses from strategic security participants or tactical security participants as to intuition level, research effort, or research time.

4.6.7 Survey 13 Result Details

The question for Survey 13 was: "How many years until Ransomware authors will send targeted deep fakes to ransomware targets? Further Details: Recipients will see realistic videos of themselves in compromising situations and will likely pay the ransom demand in order to avoid the threat of the video being released into the public domain. (please form your answer as a number and indicate years)?" This was a technical question in nature, with a presumption that technical security experts, who track adversarial techniques used by relevant threat actors and campaigns to appropriately advise their management team, would have distinct responses. As noted in Table 4.2 and Figure 4.10, there were 56 participants in this survey, and using the

Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0, which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .92, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .72, indicating that the forecasts of the two groups were not significantly different.

Given this particular technical question, I concluded that there was no significant distinction between participants. Like previous survey questions findings, this finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

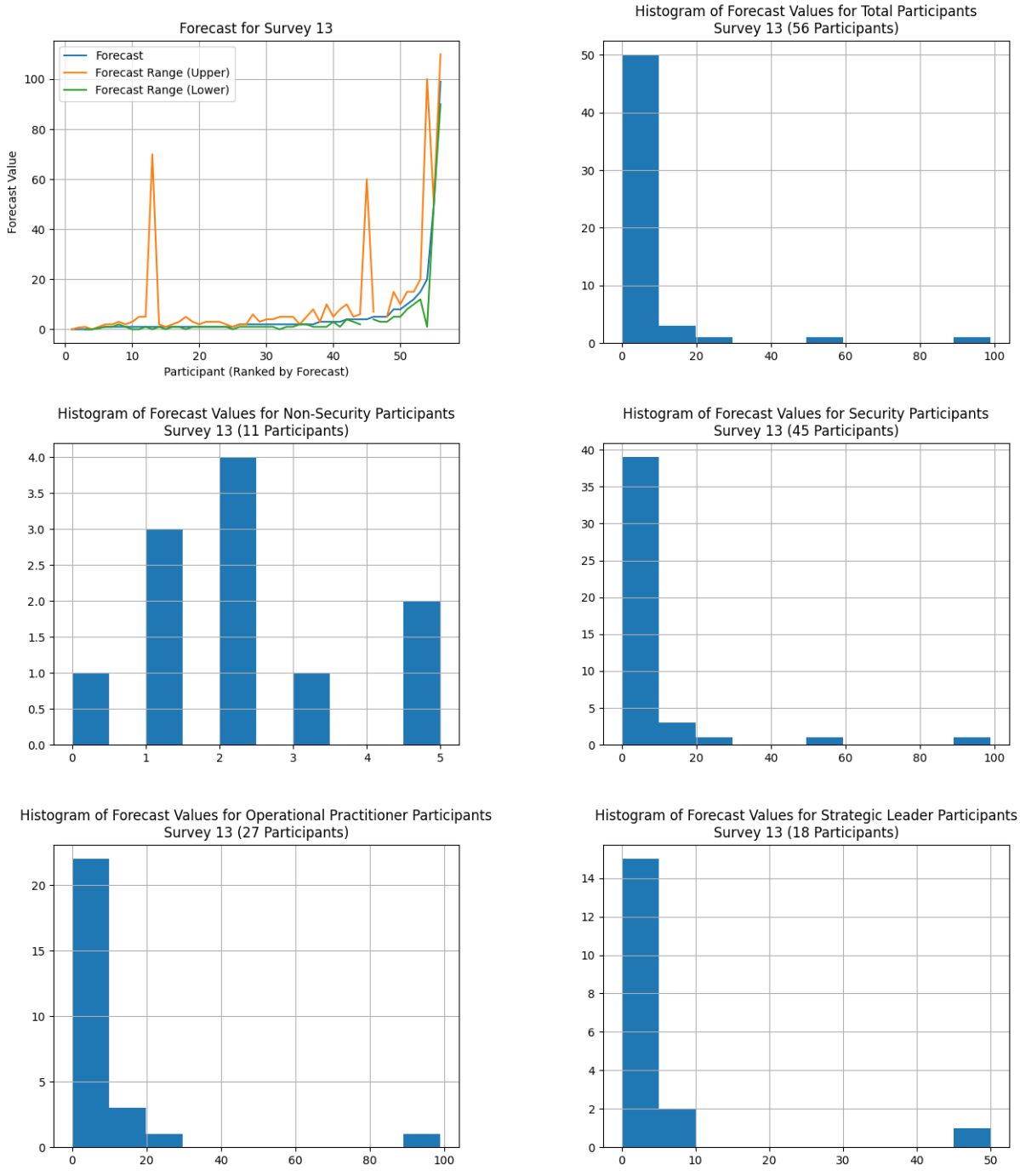

Figure 4.10: Forecast Results and Histograms of Forecast Results in Survey 13

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, using the Mann-Whitney non-parametric test, the P-Value between the security participants and non-security participants was .02 for intuition level, 0.75 for research effort, and .01 for research time. As a P-Value > 0.05 indicates a lack of difference between the responses between the two groups, this question's results demonstrated that there was a mix of distinctiveness between non-security and security participants in terms of intuition level, research effort, or research time. Specifically, the P-Value for participants' research effort showed a clear distinction between non-security and security participants.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .12 for intuition level, 0.95 for research effort, and .34 for research time. A P-Value > 0.05 indicates that we fail to find evidence of different results between the two groups. So for this particular question, the results demonstrated that strategic security participants and tactical security participants were not unique in intuition level, research effort, or research time.

4.6.8 Survey 14 Result Details

Survey 14 asked participants: "How many years until the United States (U.S.) passes a Federal Law similar to the California Consumer Privacy Act (CCPA) or the General Data Protection Regulation (GDPR)?" This was a strategic question in nature, with a presumption that strategic security experts, who track regulatory compliance law on a domestic and global level in order to plan and make business level decisions, would have distinct responses. As noted in Table 4.2 and Figure 4.11, there were 52 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0, which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .16, which is a P-Value > 0.05, indicating that the predictions of the two groups did not differ. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .22, indicating that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular strategic question, there was no significant distinction between participants. Once again, this finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

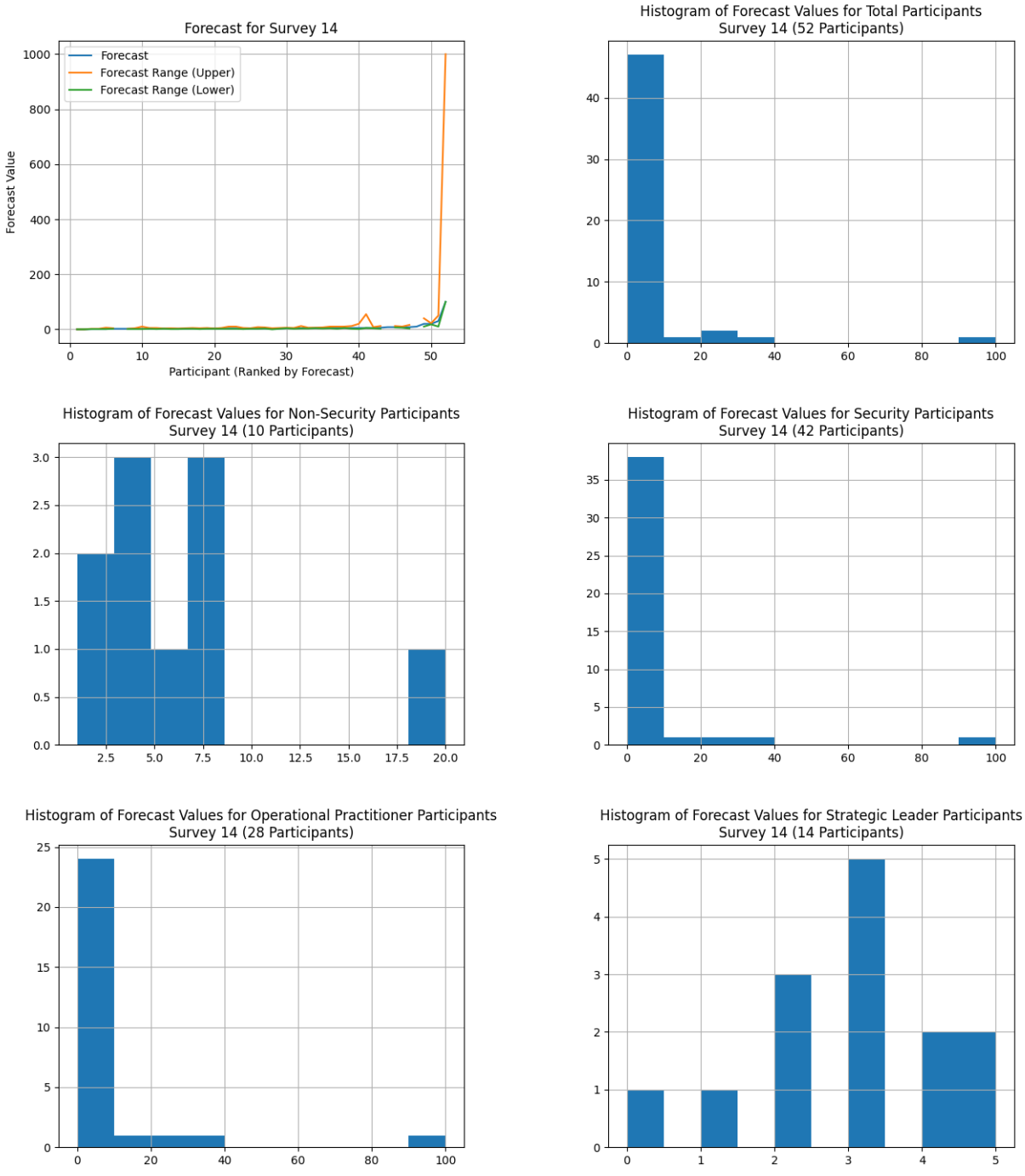

Figure 4.11: Forecast Results and Histograms of Forecast Results in Survey 14

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .47 for intuition level, 0.45 for research effort, and .04 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was a mix of distinctiveness between non-security and security participants with regards to intuition level, research effort, or research time. Specifically, the research time showed a distinction between non-security and security participants.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .37 for intuition level, 0.14 for research effort, and .01 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was a mix of distinctiveness between strategic security participants and tactical security participants with regards to intuition level, research effort, or research time. Specifically, the research time showed a distinction between strategic security participants and tactical security participants in their responses.

4.6.9 Survey 15 Result Details

In Survey 15, participants were asked: "How many months from now until we see the next CVE for a Microsoft Office 365 Vulnerability?" As this was a technical question in nature, it was presumed that technical security experts already track the occurrences of vulnerabilities within Enterprise software like Microsoft Office 365 and would have distinct responses. As noted in Table 4.2 and Figure 4.12, there were 39 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test,

indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .87, which as a P-Value > 0.05 indicates no evidence of difference between the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .4, indicating that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular technical question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

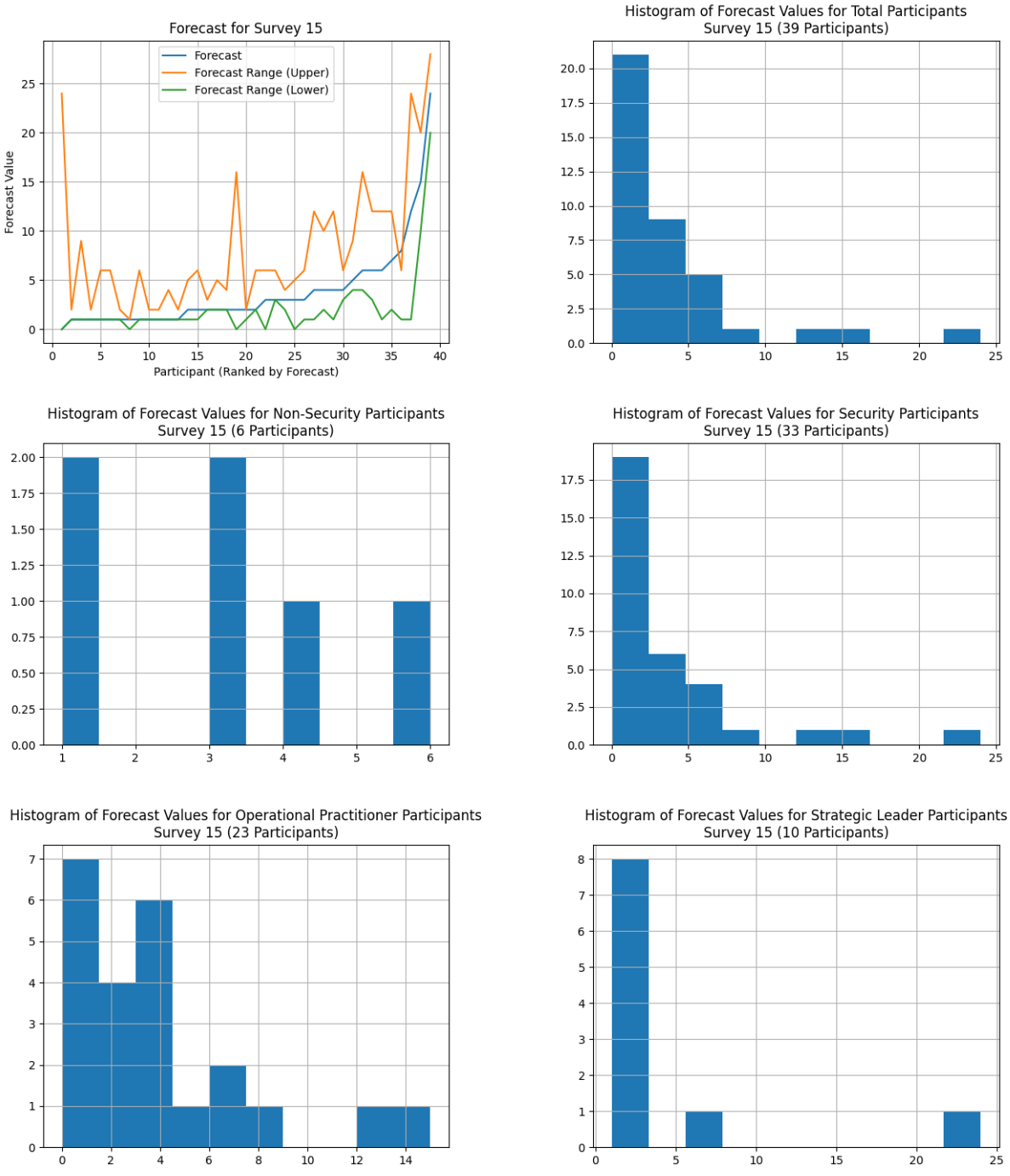

Figure 4.12: Forecast Results and Histograms of Forecast Results in Survey 15

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .01 for intuition level, 0.18 for research effort, and .06 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. For Survey 15, there was a mix of distinctiveness between non-security and security participants with regards to intuition level, research effort, or research time. Specifically, the non-security and security participants' intuition level and research time differed. The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .45 for intuition level, 0.43 for research effort, and .95 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was no unique distinctiveness between strategic security participants and tactical security participants with regards to intuition level, research effort, or research time.

4.6.10 Survey 16 Result Details

Survey 16 asked: "How many Fortune 1000 organizations will suffer a data breach in 2021 and fire or lose their CEO, CIO, or CISO within 12 months of suffering or publicly disclosing the data breach?" This was a strategic question in nature, with a presumption that strategic security experts, many who hold an executive or management title, might have experienced hiring and firing consequences due to a data breach and have distinct responses. As noted in Table 4.2 and Figure 4.13, there were 35 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .89, which is a P-Value > 0.05, indicating no evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was 1.0, which suggests that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular strategic question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

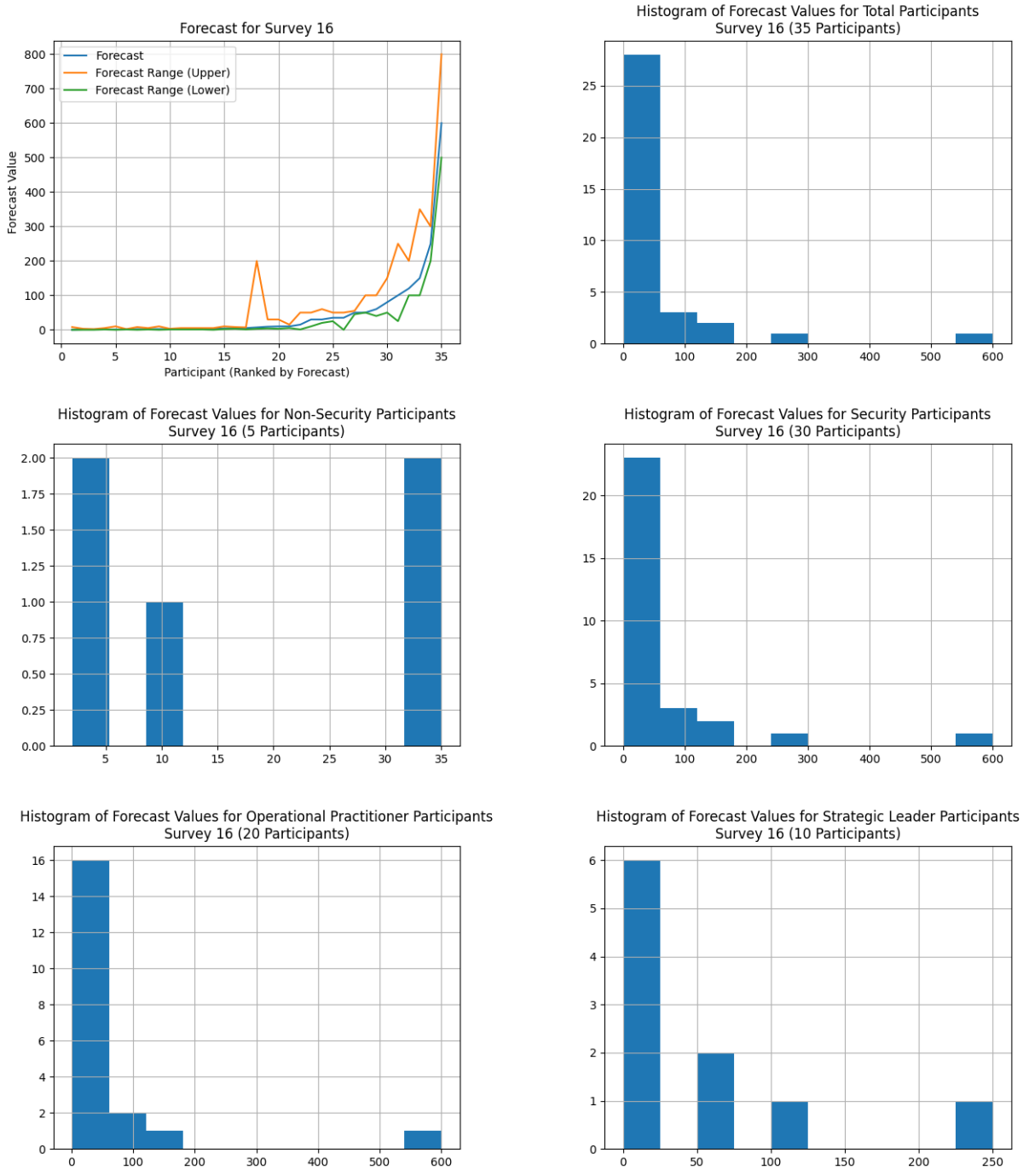

Figure 4.13: Forecast Results and Histograms of Forecast Results in Survey 16

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .02 for intuition level, 0.67 for research effort and .61 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was a mix of distinctiveness between non-security and security participants with regards to intuition level, research effort or research time. Specifically, the intuition level showed a distinction between non-security and security participants.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .74 for intuition level, 0.76 for research effort, and .96 for research time. As a P-Value > 0.05 indicates no evidence of a difference in results between the two groups, the results demonstrated that there was no distinct difference between strategic security participants' and tactical security participants' intuition levels, research efforts, and research time.

4.6.11 Survey 17 Result Details

The question for Survey 17 was: "What is the percentage likelihood that we will read publicly about attackers exploiting infrastructure that will lead to a DDoS or DoS condition affecting the 2020 U.S. National Elections?" This was a strategic question in nature, with a presumption that strategic security experts track public data breaches and attacks, specifically related to National elections, and would provide distinct responses. As noted in Table 4.2 and Figure 4.14, 43 participants responded to the question, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0.000549, which failed the normality test and reveals that the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .48, which is a P-Value > 0.05 and shows that there is no difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .53, indicating that the forecasts of the two groups were not significantly different.

From these results, I conclude that there was no significant distinction between participants when answering this particular strategic question. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

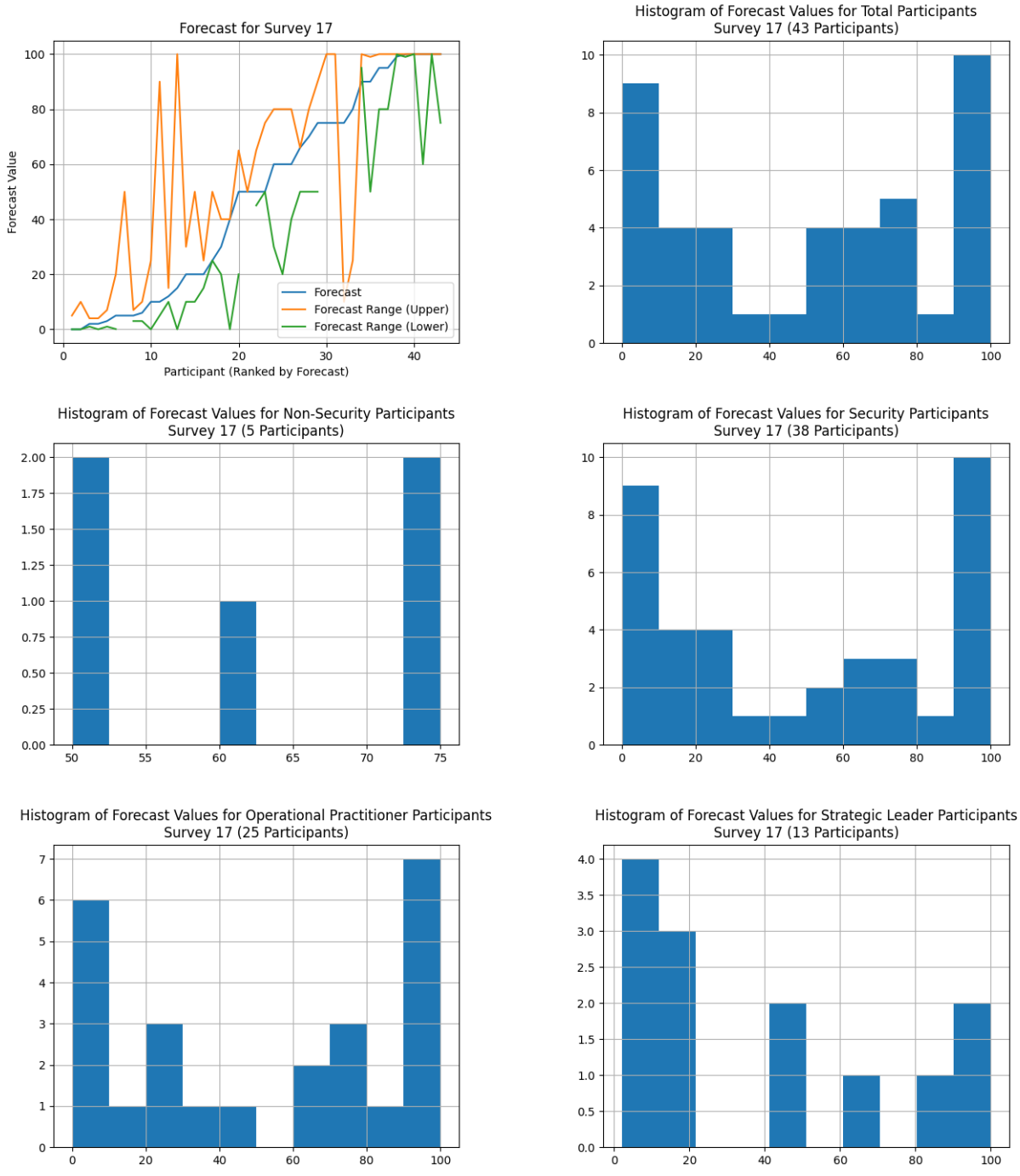

Figure 4.14: Forecast Results and Histograms of Forecast Results in Survey 17

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .14 for intuition level, 0.5 for research effort, and .18 for research time. If a P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups, then Survey 17's results demonstrated that there was no unique distinctiveness between non-security and security participants with regards to intuition level, research effort, or research time. The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .33 for intuition level, 0.68 for research effort, and .95 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was no unique distinctiveness between strategic security participants and tactical security participants with regards to intuition level, research effort, or research time.

4.6.12 Survey 18 Result Details

Survey 18 asked participants: "How many years until a typical cybersecurity budget for an organization surpasses 20% or greater of the total IT budget?" This was a strategic question in nature, with a presumption that strategic security experts, many who manage the security budget, would have distinct responses. As noted in Table 4.2 and Figure 4.15, there were 32 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .55, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .12, indicating that the forecasts of the two groups were not significantly different, which was contrary to my assumption that there would be significant difference.

I therefore conclude that given this particular strategic question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

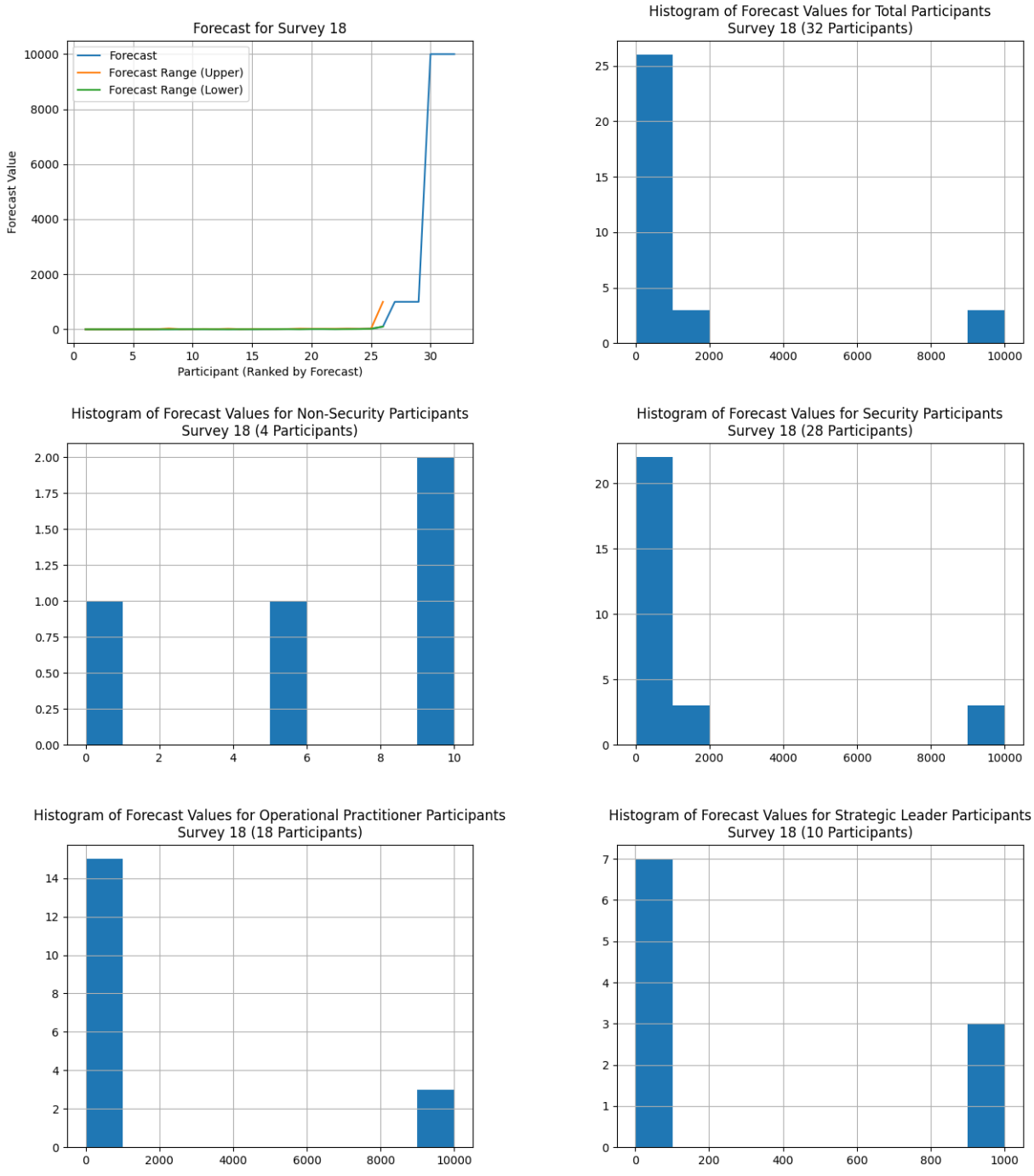

Figure 4.15: Forecast Results and Histograms of Forecast Results in Survey 18

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was 0.0 for intuition level, 0.75 for research effort, and .13 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. For Survey 18n, the results demonstrated that there was a mix of distinctiveness between non-security and security participants with regards to intuition level, research effort or research time. Specifically, the intuition level showed a distinction between non-security and security participants.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .94 for intuition level, 0.73 for research effort and .4 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was no unique distinctiveness between strategic security participants and tactical security participants with regards to intuition level, research effort or research time.

4.6.13 Survey 19 Result Details

In Survey 19, participants were asked: "Recent vulnerabilities were announced AKA Achilles vulnerabilities, detailing how flaws in Qualcomm Snapdragon chips could be exploited to monitor location and audio and to steal images and videos. They could also be exploited to render devices useless. The chips are used in hundreds of millions of Android devices. What is your percentage confidence that there will be a public story of these vulnerabilities (AKA Achilles CVEs) being publicly exploited before fixes are incorporated into the Android OS or Android devices that use Snapdragon?" A technical question in nature, I assumed that technical security experts, who track exploitation likelihood take action and communicate to management

recommendations related to security resource and project prioritization, would have distinct responses. As noted in Table 4.2 and Figure 4.16, there were 24 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that the dataset had a P-Value of 0.011228 which in this case meet the criteria required of the normality test, indicating the dataset did follow a normal distribution ($p \leq 0.05$).

Forecast Results

Although the dataset followed normal distribution, the Mann-Whitney non-parametric test was still used to perform significance tests between data sets as it is an even more relevant and powerful test for datasets that follow a normal distribution. The Mann-Whitney P-Value between the security participants and non-security participants was .28, which is a P-Value $>$ 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .53, indicating that the forecasts of the two groups were not significantly different.

Therefore, it can be reasonably concluded that there was no significant distinction between participants' forecasts for this technical question. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

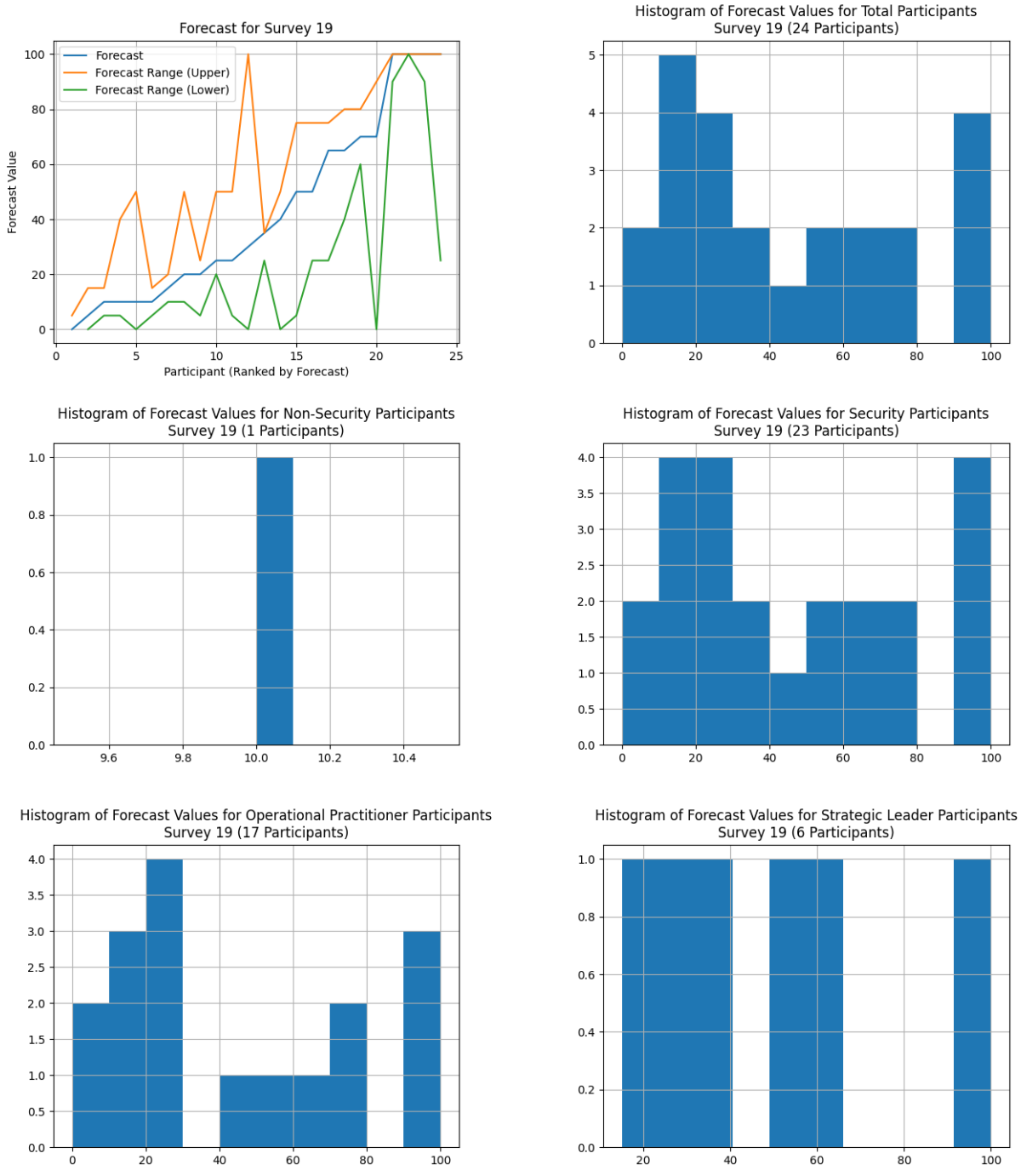

Figure 4.16: Forecast Results and Histograms of Forecast Results in Survey 19

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .06 for intuition level, 0.1 for research effort, and .1 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. For this particular question, the results demonstrated that there was no unique responses from non-security and security participants with regards to their assessment of their own intuition level, research effort, or research time.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .36 for intuition level, 0.02 for research effort, and .8 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. The results of this particular question demonstrated that there was a mix of distinctiveness between security strategy and technical participants with regards to their intuition levels, research effort, or research time. Specifically, strategic security participants' and tactical security participants' level of intuition differed.

4.6.14 Survey 20 Result Details

The forecasting question for Survey 20 was: "What percentage of Fortune 1000 companies will have a documented response plan for pandemics as part of their business continuity and disaster recovery strategy within 12 months from now?" This was a strategic question in nature, with a presumption that strategic security experts have the expertise and the necessary background related to understanding how an event like the COVID-19 pandemic, which was very relevant at the time of asking the question, would affect an organization's response plan and that these qualifications would result in distinct responses. As noted in Table 4.2 and Figure 4.17, there were 36 participants in this survey, and using the Shapiro-Wilk normality tests, it was shown that

the dataset had a P-Value of 0.006135 which failed the normality test, indicating the dataset did not follow a normal distribution ($p \leq 0.05$).

Forecast Results

Because the dataset did not meet the criteria for normal distribution, the Mann-Whitney non-parametric test was used to compare the forecasts given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .93, which is a P-Value > 0.05, indicating that we fail to find evidence of a difference in the predictions of the two groups. Similarly, the Mann-Whitney P-Value between the strategic security participants and tactical security participants was .08, indicating that the forecasts of the two groups were not significantly different.

I therefore conclude that given this particular strategic question, there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant.

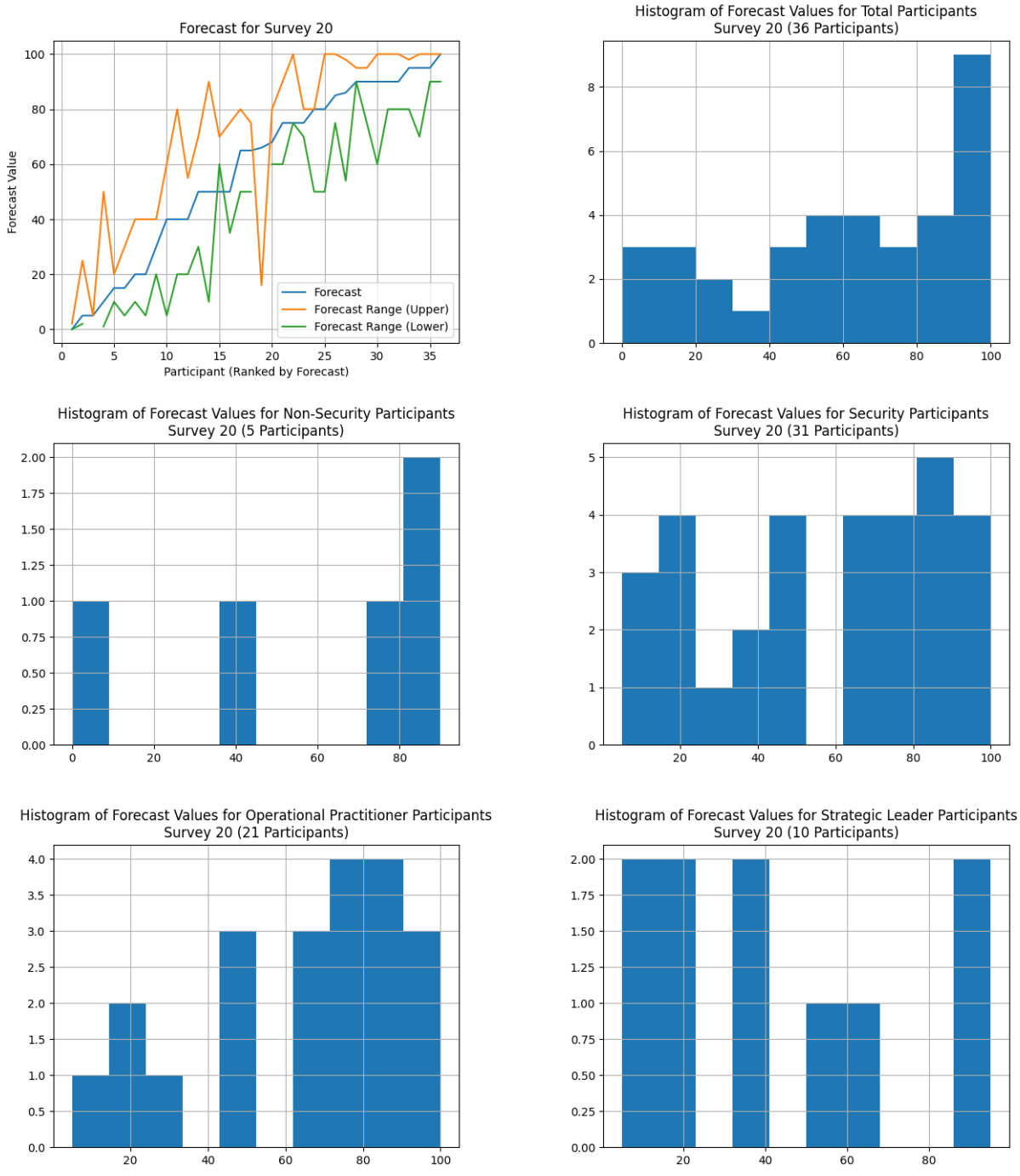

Figure 4.17: Forecast Results and Histograms of Forecast Results in Survey 20

Intuition Level, Research Effort, and Research Time Results

As noted in Tables 4.3, 4.4, and 4.5, the Mann-Whitney non-parametric test was used to compare the responses given by the different groups. The Mann-Whitney P-Value between the security participants and non-security participants was .22 for intuition level, 0.37 for research effort, and .21 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was no unique distinctiveness between non-security and security participants with regards to intuition level, research effort, or research time.

The Mann-Whitney P-Value between the strategic security participants and tactical security participants was .82 for intuition level, 0.07 for research effort, and .17 for research time. A P-Value > 0.05 indicates that we fail to find evidence of a difference in results between the two groups. So for this particular question, the results demonstrated that there was a mix of distinctiveness between security strategy and technical participants with regards to intuition level, research effort, and research time. Specifically, the P-Value of research effort showed a clear distinction between strategic security participants and tactical security participants.

4.7 Discussion

From the twenty surveys that were conducted, Surveys 7-20 were considered for analysis, although only Surveys 11-20 had sufficient data from which to form conclusions. While some conclusions were gathered from the results, the level of participation was dissatisfying. In general, motivating participants to respond in survey type studies was difficult, even after direct follow-up reminders. As the majority of the data could not be considered "Gaussian," or pass normality tests, non-parametric testing was used and the resulting analysis found that there was no significant distinction between participants. This finding is contrary to the traditional belief that the prediction given by a security participant would be distinctive from that given by a non-security participant, which is contrary to the initial assumptions I made at the beginning of this study.

Chapter 5

Limitations and Challenges

5.1 Limitations

The most obvious limitation to this study was that the security expert participants were solicited by the author's network versus soliciting a larger, broader audience. The author has been in security for over 20 years, so the qualifications of the security experts are not in question, but rather bias may exist because the experts were selected through the network of one individual and may consist of too narrow a type of individual in technical fields. Additionally, the survey questions were selected by the author and perhaps different questions would give different answers.

5.2 Challenges

A number of challenges existed during the course of this research study. First, generating questions that were relevant, timely, and interesting and reflected questions related to the decisions leaders were asked to make proved to be difficult and a significant task of its own right. Second, and most notably, soliciting responses from participants proved to be incredibly challenging and

inconsistent. Overall, the surveys that were analyzed had a response rate of 38%, which was far lower than what was predicted. Although over 150 participants had willingly committed to participation, in reality, we typically had barely enough participation to gather reliable data from which to draw significant conclusions.

Chapter 6

Conclusion

6.1 Background on Expert Opinion in Cybersecurity

In today's cybersecurity industry, security experts are increasingly relied on to provide judgments in order for leaders to make optimal decisions. These experts are assumed to provide distinct judgments over asking a general technical operator. Whereas past literature has questioned the accuracy, acceptance, and rejection of such expert advice and judgments, this research study focuses on whether there exist patterns of significant difference between the forecasts of security professionals vs non-security technical professionals. If we can not find a significant difference, it could indicate that security decisions today, many of which are founded on security expert opinions, have no more value than asking a non-security technical individual.

6.2 Survey Results and Contributions

The initial industry assumption this thesis was focused on was the notion that security experts provide a distinct and more accurate judgment compared to non-security participants when helping to forecast decisions on cybersecurity topics. Using a method of regular surveys

that were sent to both security experts and non-security participants over a number of months, the data indicates that there was no significant difference or unique attribute to the judgments made by security experts vs non-security participants. Moreover, this proved to be true for questions regarding both strategic topics and technical topics. What this would lead me to conclude is that I can not support the claim that security professionals offer a distinct judgment over non-security experts.

6.3 Further Research

Given the aforementioned limitations and challenges of this work, the following areas below could be of interest for further investigation.

6.3.1 Increasing the Number of Participants in the Study

In this research study, 150 individuals participated in surveys over a number of months. The security experts were hand-selected by the author and the non-security participants were referred to us and then partially hand-selected. As mentioned before, we had a 38% response rate, and when we broke up the participants into the associated datasets and labeled cohorts, the number of participants was in many cases lower than necessary for certain algorithms to produce significant conclusions from. To optimize for a greater number of responses, two improvements that could be made for further research would be to increase the number of participants surveyed and to consider offering compensation to increase participation. As most participants are professionals, who receive a salary, compensation could be something other than money.

6.3.2 Introducing and Measuring the Effect of Forecast Training

As was discussed in Chapter 2, the field of forecast judgment has matured greatly in the last 10 years. Projects like the Good Judgement Project (GJP) have exposed how individual characteristics can result in more accurate forecasting but have also published work on forecast training so that both individuals and groups can become better at forecasting accurately. The subject of forecast training is rarely discussed amongst cybersecurity professionals and it would be valuable to expand on this research study by soliciting a research group made up of security experts, training half of those experts in better forecasting techniques as outlined by groups like the GJP and measuring whether the training provided more accurate and valuable judgments.

6.3.3 Measuring Individual Responses Versus Group Responses

As the GJP indicated, individuals are often highlighted as the sole determinants of cybersecurity forecasts. Typically, when we read a quote or advice in the news related to cybersecurity, it is attributed to a single individual person. It is also common that within a security team or consulting group, there is a more senior individual who is asked to provide judgments so that decisions can be made. Yet, more research is needed to understand the different roles and contributions individuals and groups make when forecasting. A potential and valuable extension of this research would examine that question. For example, a research group could be formed and split into two groups. One group would be asked to forecast judgments as individuals and the other group broken into groups of 3 or more individuals and asked to make judgments as a group. In analyzing the results, it would be valuable to determine if group judgments both provide unique perspectives as well as more accuracy compared to the individual forecasters.

Bibliography

[1] L. Bilge, Y. Han, and M. Dell'Amico. RiskTeller: Predicting the Risk of Cyber Incidents. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS), Dallas, Texas*, 2017.

[2] D. Canali, L. Bilge, and D. Balzarotti. On the Effectiveness of Risk Prediction Based on Users Browsing Behavior. *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (CCS), Kyoto, Japan*, 2014.

[3] A. Chhabra, P. Iyengar, and J. Lopez. Toolkit: Presentation for Key Findings From the 2020 Board of Directors Survey. https://www.gartner.com/en/documents/3976147/toolkit-presentation-for-key-findings-from-the-2020-boar, 2019. Accessed: 2019-12-10.

[4] Review of Issues Related to the Loss of VA Information Involving the Identity of Millions of Veterans. https://www.va.gov/oig/pubs/VAOIG-06-02238-163.pdf, 2006.

[5] Cybersecurity Statistics. https://www.fortinet.com/resources/cyberglossary/cybersecurity-statistics, 2021.

[6] J. Frost. One-Tailed and Two-Tailed Hypothesis Tests Explained. https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/, 2021.

[7] Good Judgement Open Crowd Forecasting Site. https://gjopen.com, 2015.

[8] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. *NSPW '09, Oxford, England*, page 133–144, 2009.

[9] A. Hernández. A Walk Through Historical Correlations Between Vulnerabilities Stock Prices. Presented as the 2021 Black Hat Conference, Asia, 2021.

[10] M. Horowitz. Good judgment in forecasting international affairs (and an invitation for season 3). https://www.washingtonpost.com/news/monkey-cage/wp/2013/11/26/good-judgment-in-forecasting-international-affairs-and-an-invitation-for-season-3/, 2013.

[11] Aggregative Contingent Estimation (ACE). https://www.iarpa.gov/index.php/research-programs/ace. Accessed: 2014-05-06.

[12] I. Ion, R. Reeder, and S. Consolvo. "...No one Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. *Presented at Symposium on Usable Privacy and Security, Ottawa, Canada*, 2015.

[13] M. Lawrence, P. Goodwin, M. O'Connor, and D. Önkal. Judgmental forecasting: A review of progress over the last 25 years. *Journal of Human Evolution*, 22(3):493–518, 2006.

[14] K. McCoy. Target to pay $18.5M for 2013 data breach that affected 41 million consumers. https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/, 2017.

[15] N. M. Razali and Y. B. Wah. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[16] E. M. Redmiles, S. Kross, and M. L. Mazurek. How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. *CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria*, page 666–677, 2016.

[17] R. Reeder, I. Ion, and S. Consolvo. 152 Simple Steps to Stay Safe Online: Security Advice for Non-tech-savvy Users. *IEEE Security and Privacy*, 15(5):55–64, June 2017.

[18] H. Smith. Hospital System Sued After Significant Cybersecurity Breach. https://www.law.com/insurance-coverage-law-center/2021/07/14/hospital-system-sued-after-significant-cybersecurity-breach/, 2021.

[19] R. Sobers. 134 Cybersecurity Statistics and Trends for 2021. https://www.varonis.com/blog/cybersecurity-statistics/, 2021. Accessed: 2021-03-16.

[20] A. Spiegel. So You Think You're Smarter Than A CIA Agent. https://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent, 2013. Accessed: 2014-07-18.

[21] M. Sponauer. The Ten Worst Tech Predictions of All Time. https://www.laptopmag.com/articles/10-worst-tech-predictions-of-all-time, 2013.

[22] D. Swinhoe. 7 security incidents that cost CISOs their jobs. https://www.csoonline.com/article/3510640/7-security-incidents-that-cost-cisos-their-jobs.html, 2020.

[23] P. Tetlock. *Expert political judgment: How good is it? How can we know?* N.J.: Princeton University Press, 2006.

[24] P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015. Available: Amazon Kindle.

[25] L. Ungar, B. Mellors, V. Satopää, J. Baron, P. Tetlock, J. Ramos, and S. Swift. The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. *AAAI Fall Symposium Series*, 2012.

[26] The Global Risks Report 2019. https://www.weforum.org/reports/the-global-risks-report-2019, 2019.

[27] B. W. Yap and C. H. Sim. Comparison of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.