

**IS THE GLASS HALF EMPTY OR HALF FULL?  
AN EXPERIMENTAL STUDY OF BAYESIAN VERSUS FREQUENTIST  
STATISTICS' INFLUENCE ON PROGRAM ENDORSEMENTS BY  
LEGISLATIVE STAFF**

by

Andrew Hurwitz

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education

Summer 2020

© 2020 Andrew Hurwitz  
All Rights Reserved

**IS THE GLASS HALF EMPTY OR HALF FULL?  
AN EXPERIMENTAL STUDY OF BAYESIAN VERSUS FREQUENTIST  
STATISTICS' INFLUENCE ON PROGRAM ENDORSEMENTS BY  
LEGISLATIVE STAFF**

by

Andrew Hurwitz

Approved: \_\_\_\_\_  
Chrystalla Mouza, Ed.D.  
Director of the School of Education

Approved: \_\_\_\_\_  
Gary T. Henry, Ph.D.  
Dean of the College of Education and Human Development

Approved: \_\_\_\_\_  
Douglas J. Doren, Ph.D.  
Interim Vice Provost for Graduate and Professional Education and  
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Henry May, Ph.D.  
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Douglas Archbald, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Kelley Borradaile, Ph.D.  
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

---

Joshua Wilson, Ph.D.  
Member of dissertation committee

## ACKNOWLEDGMENTS

I started my research journey when I was young. Part of me was trying to keep up with my overachieving high school classmates and the other part of me curious about what science was like outside of high school. As such, I took advantage of an opportunity at the age of 15 to join a toxicology lab at the University of Sciences in Philadelphia (USP). Yes, I have my father to thank for making that position happen, and since this is an acknowledgements section, I owe a special thank you to him. He had an appointment in the Department of Clinical Pharmacy and at the time, risked his reputation by having me participate in that lab. I am not sure I deserved that confidence during that time in my academic journey but nonetheless, I hope my accomplishments there made him proud.

In that lab, I participated in experiments that sought to understand how chemical solutions ultimately used in contact lens could reduce irritation to the human cornea. I remember little about those experiments now, other than the principal investigator, Dr. Ruy Tchao, went on to win many patents for contact lens solution, so it would be my assumption that the experiments were a success.

After graduating from high school, I became more interested in humans than cells, and I spent three years using qualitative methods to study scientific reasoning and educational strategies for promoting conceptual change about evolution. My undergraduate thesis examined epistemological and ontological worldviews and how those worldviews affected scientific reasoning. My thesis was presented at the American Educational Research Association, which was my first exposure to the

professionalization of research. Ironically enough, that study also was a vignette project.

After earning my bachelor's degree, I spent a year as a lab manager at a medical school bridging the biological and social sciences where I used functional and structural magnetic resonance imaging to assess brain changes in pediatric populations diagnosed with autism. Frustrated with the small sample sizes of University studies, I left for Mathematica where my research has spanned from randomized control trials of alternative styles in tackle football to developing intuitive and real-time data analytic dashboards that enable non-statistician audiences to easily digest descriptive and inferential information regarding their data assets. What has been central to my work is methods, and I was fortunate to find my home at the University of Delaware studying evaluation, measurement, and statistics in education.

I have been part of various research communities for more than half my life now, and these opportunities would not have presented themselves without my parents, Tammy and Ben Hurwitz. My parents are humble individuals who worked first after graduating high school and started at community college before each eventually earning their bachelor's degree. My father would go on to become a pharmacist and later earn an adjunct appointment at USP. I loved spending the last weeks of August on campus with my father as he taught students and mentored the next generation of pharmacists. We had our own dialogue to break the ice on the first day of class. I would play the boy genius who knew all the answers to the questions. I still remember those days with great fun and affection. My mother left a successful career in real estate to focus on me and my brother during our formative years. She returned to college after beating breast cancer and achieved her life dream by

becoming a nurse at the age of 62. For the past four years she has provided outstanding care throughout South Jersey to numerous pediatric patients who fight every day through life threatening conditions. Mom, I love you, and you are my hero. My parents taught me to ask questions, challenge assumptions, advocate for myself, plan for my future, and work tirelessly to achieve my goals. My appreciation for their love, encouragement, and endearing sense of pride in their boy is something words can never fully capture.

I also want to thank my brother, Michael. He has taught me patience, self-reflection, and the true love that comes with a brother. I know when I need him, he is there for me, and I know he feels the same. My wife Felicia has been my partner in crime since high school. Completing our doctoral education together has meant everything to me. Many people told us we should stagger our studies, but we knew that to do this well meant completing it together. We have been through much in our lives that has caused old souls in young people, and her love and support keep me grounded and motivated when obstacles seem impossible to overcome. Words also cannot capture how much she has done for me throughout the years and these sentences do not even come close to the mark on how lucky I am to have her in my life. Out of all the titles I have earned and will earn in my life, the one that I will always feel most privileged to have is her husband.

Third, I wish to thank my Cockapoo, Tucker. For those that are fortunate to call a dog their friend, the bond is understood. Fourth, I wish to thank my extended family and friends. Their support has meant the world these past five years, and it is my sincere hope that I support them as much as they have made me feel supported.

Finally, I wish to thank my doctoral advisor, Dr. Henry May and doctoral committee, Dr. Douglas Archbald, Dr. Kelley Borradaile, and Dr. Joshua Wilson. Each member of my doctoral committee has played a true special role in my professional development. Dr. Archbald was an early mentor in graduate school, and he supported me in expanding my interests beyond my early work at Delaware allowing me to connect the policy world of Mathematica to my graduate studies. Dr. Borradaile watched my professional career blossom from my early days at Mathematica when she was my supervisor, to helping me develop my leadership skills on our concussion project, and now as a member of my dissertation committee. Her words of encouragement and support always made me feel like I had somebody who understood, and I am forever grateful. It is my sincere hope that I can do for another what she has done for me. Dr. Wilson and I have worked together less often, but I recall fondly a deep intellectual conversation we had during my first year in the program regarding my work at Mathematica in natural language processing and his research on strategies for supporting student writing. My dissertation has benefited immensely from his contributions, and I am forever grateful for his willingness to lend his support and time to my professional development.

Finally, I wish to extend a special thanks to Dr. Henry May. I have completed numerous classes in advanced statistics and research methods with Dr. May and in addition, Dr. May has supervised several of my independent studies. My first Association for Public Policy Analysis and Management conference was to present work conducted under Dr. May. His commitment to students is one of a kind. I remember when I first approached Dr. May about this dissertation idea. To conduct this work meant having to switch my specialization, and Dr. May embraced these

implications welcoming me as his graduate student and empowering me to conduct meaningful graduate work that truly reflected my scholarly interests. I know without a doubt, that there is one more PhD in this world because of Dr. May. It is my firm belief that without Dr. May's support and mentorship, I would not have completed my doctoral program.

## TABLE OF CONTENTS

LIST OF TABLES .....	xii
LIST OF FIGURES .....	xiv
ABSTRACT .....	xv
Chapter	
1 INTRODUCTION .....	1
Bayesian and frequentist methods: An evolving landscape .....	2
Core differences between the Bayesian and frequentist approaches .....	5
Why care about differences between Bayesian and frequentist statistics? .....	6
Early empirical evidence on Bayesian decision making .....	9
The case for studying statistical judgment .....	10
High stakes decision making: More than just evidence .....	12
2 LITERATURE REVIEW .....	16
A Very Brief Tale of Two Statistical Paradigms .....	16
Hypothesis Testing for Inference in Experiments .....	16
Concepts Relevant to Decision Making for Statistical Inference .....	18
P-Values, Power, and Posterior Probabilities .....	18
Concepts Relevant to Precision for Statistical Inference .....	23
Standard Errors, Confidence Intervals, and Credible Intervals .....	23
Additional Concepts Relevant to Decision Making for Statistical Inference .....	26
Effect Sizes, Tests of Equivalence, and Regions of Practical Interest .....	26
Revisiting Core Differences: Bayesian v. Frequentist .....	28
Infants (and Adults) as Natural Bayesians .....	34
A Final Word on Training to be Bayesians .....	35
Statistical Judgement Model (SJM): Bayesian versus Frequentist .....	37
3 METHODS .....	42
Overview .....	42
4 RESULTS .....	52

	Descriptive Statistics .....	53
	Congressional Staffers.....	53
	University of Delaware Undergraduates .....	56
	Baseline Tests of Equivalency.....	59
	Congressional Staffers.....	59
	University of Delaware Undergraduates .....	60
	Frequentist ANOVA Models.....	61
	Congressional Staffers.....	62
	University of Delaware Undergraduates .....	64
	Bayesian ANOVA Model.....	66
	Congressional Staffers.....	68
	University of Delaware Undergraduates .....	69
5	DISCUSSION.....	70
	Contributions to Research .....	72
	Implications for the Policy Research Community.....	74
	Implications for Congressional Staffers .....	76
	Implications for University Instructors Teaching Statistics .....	77
	Beyond Cognitive Science: Fundamental Ontological Differences .....	78
	Limitations.....	81
6	CONCLUSION .....	83
	REFERENCES .....	87
	Appendix	
A	SAMPLE EMAIL FOR RECRUITMENT OF LEGISLATIVE AIDES.....	94
B	SAMPLE EMAIL TO INSTRUCTORS OF POLITICAL SCIENCE UNDERGRADUATES .....	95
C	SAMPLE RECRUITMENT SPEECH TO POLITICAL SCIENCE UNDERGRADUATES .....	96
D	DATA COLLECTION TIMELINE.....	97
E	VIGNETTES .....	98
	Sample Vignette Frequentist: Moderate Evidence + Low Cost + Difficult Implementation .....	98
	Sample Vignette Bayesian: Moderate Evidence + Low Cost + Difficult Implementation .....	99
	Sample Vignette Frequentist: Moderate Evidence + High Cost + Easy Implementation .....	100
F	VIGNETTE QUESTIONNAIRE .....	102

G	FREQUENTIST LOGISTIC REGRESSION MODELS .....	111
	Congressional Staffers .....	111
	University of Delaware Undergraduates .....	113
H	CUMULATIVE RESPONSES .....	116
	Congressional Staffer Graphs .....	117
	Undergraduate Graphs .....	123
I	FREQUENTIST ANOVA MODELS FOR POOLED SAMPLE .....	129
	Pooled Sample of Congressional Staffers and Undergraduates .....	129
J	BENJAMINI-HOCHBERG PROCEDURE .....	132
K	IRB APPROVAL LETTER .....	133

## LIST OF TABLES

Table 1	Reported Age of Congressional Staffers .....	54
Table 2	Reported Undergraduate Majors of Congressional Staffers.....	54
Table 3	Reported Statistics Courses Completed by Congressional Staffers .....	55
Table 4	Reported Public Policy Work Area for Congressional Staffers .....	56
Table 5	Reported Year of Study for Undergraduates .....	57
Table 6	Reported Second Undergraduate Major for Undergraduates .....	57
Table 7	Reported Statistics Courses Completed by Undergraduates .....	58
Table 8	Reported Public Policy Work Area of Interest for Undergraduates .....	59
Table 9	Baseline Equivalency Tests for Age and Undergraduate Major for Congressional Staffers.....	60
Table 10	Baseline Equivalency Tests for Statistics Courses and Policy Area for Congressional Staffers.....	60
Table 11	Baseline Equivalency Tests for Age and Second Major for Undergraduates.....	61
Table 12	Baseline Equivalency Tests for Statistics Courses and Policy Area for Undergraduates.....	61
Table 13	Overall Frequentist ANOVA Values for Each of the Items by Paradigm for the Congressional Staffers .....	64
Table 14	Overall Frequentist ANOVA Values for Each of the Items by Paradigm for the Undergraduates .....	66
Table 15	Overall Bayesian ANOVA Values for Each of the Models for the Congressional Staffers.....	68
Table 16	Overall Bayesian ANOVA Values for Each of the Models for the Undergraduates.....	69
Table 17	Overall Frequentist Repeated Measures Logistic Regression Values for Each of the Items by Paradigm for the Congressional Staffers .....	113

Table 18	Overall Frequentist Repeated Measures Logistic Regression Values for Each of the Items by Paradigm for the Undergraduates .....	115
Table 19	Overall Frequentist ANOVA Values for Each of the Items by Group..	131
Table 20	Benjamini-Hochberg Procedure for Congressional Staffers .....	132
Table 21	Benjamini-Hochberg Procedure for Undergraduates .....	132

## LIST OF FIGURES

Figure 1	Statistical Judgment Model: Bayesian v. Frequentist.....	40
Figure 2	Presentation of Study Hypotheses by Arms Tested in this Experiment..	41
Figure 3	Cumulative Counts for Question 1 .....	117
Figure 4	Cumulative Counts for Question 2 .....	118
Figure 5	Cumulative Counts for Question 3 .....	119
Figure 6	Cumulative Counts for Question 4 .....	120
Figure 7	Cumulative Counts for Question 5 .....	121
Figure 8	Cumulative Counts for Question 6 .....	122
Figure 9	Cumulative Counts for Question 1 .....	123
Figure 10	Cumulative Counts for Question 2 .....	124
Figure 11	Cumulative Counts for Question 3 .....	125
Figure 12	Cumulative Counts for Question 4 .....	126
Figure 13	Cumulative Counts for Question 5 .....	127
Figure 14	Cumulative Counts for Question 6 .....	128

## ABSTRACT

The Bayesian and frequentist statistical paradigms are the two most well-known paradigms in the field of statistical inference. Bayesian statisticians consider inferences through increasing degrees of certainty where statistical results are presented as probability statements. On the other hand, the frequentist paradigm is best known for its use of p-values and confidence intervals to express results in terms of significance and precision. Despite the long-standing tradition of both these paradigms, the frequentist paradigm has for many years been the dominant paradigm in statistical training for social scientists. The reasons for the dominance of the frequentist paradigm are widely debated. In recent years however, the use of Bayesian methods by social scientists to analyze data has increased, including analyses intended to inform high stakes decision making. Some proponents of the Bayesian paradigm argue that probability statements from Bayesian analyses are easier to understand than confidence intervals and p-values. Recognizing that Bayesian methods are used in such high stakes fields like public policy research where the results of an impact evaluation are used to inform decisions such as whether or not to fund social programs (e.g., providing services to impoverished mothers, job training to disabled individuals, etc.), understanding the role Bayesian statistics might have in influencing statistical judgment is important.

The claim that probability statements are easier to understand than confidence intervals and p-values is one that can be tested empirically. Early research in understanding the role Bayesian versus frequentist statistics have in influencing decision making regarding whether respondents would endorse a new educational technology is already underway (Chandler, Martinez, Finucane, Terziev, & Resch,

2019). This dissertation sought to expand upon this important work and conducted a statistical vignette experiment with United States congressional aides to understand whether legislative aides are more likely to endorse an education program when results from an effectiveness evaluation of that program are presented under a Bayesian versus frequentist paradigm.

Thirty congressional aides were randomly assigned to one of four conditions. Each condition presented the same level of equivalent evidence in either the frequentist or Bayesian paradigm. Information regarding the cost of the program as well as the feasibility of implementation was incorporated into the vignettes resulting in two different cost and implementation scenarios (i.e., a low cost and difficult to implement scenario and a high cost and easy to implement scenario). Participants were asked to respond to six survey items. Two items aimed to assess whether respondents found the information presented in the vignettes as informative and easy to understand whereas the remaining four vignettes asked participants to rate their level of agreement with whether an endorsement of the program is justified. A second sample consisting of thirty-six undergraduates majoring in political science also participated in an identical vignette experiment.

Results from the experiment were analyzed using both frequentist and Bayesian factorial repeated measures ANOVA models demonstrating statistically significant findings on all six dependent variables for the Congressional sample and three models showing statistically significant results in the sample of undergraduates. Bayes factor was used to interpret the results from the Bayesian ANOVA suggesting findings consistent with the frequentist results. All effects indicated more favorable reactions to the evidence when presented under the Bayesian paradigm. Implications

for this work are discussed in terms of their application to statistical methods for social science research, and considerations for when to present results in a Bayesian versus frequentist framework are discussed. Finally, limitations of this work are addressed with respect to the need for future experiments to test other salient characteristics of statistical information in their vignettes as well as the need for replication of experimental results. Future directions for the work are suggested regarding the need to establish community consensus from researchers with respect to the best practices for presenting Bayesian statistical information to Congressional staffers.

*Keywords:* Bayesian statistics, frequentist statistics, human judgement

## **Chapter 1**

### **INTRODUCTION**

The utilization of statistical procedures for generating evidence to make informed decisions is the hallmark of contemporary scientific thought in the applied social sciences. In fields such as clinical medicine and public policy where decisions are deemed high stakes, the presence of such rigorous evidence is argued as a necessity for decision making (Montorri & Guyatt, 2008; Head, 2008). Statisticians and applied social scientists have spent decades developing guidelines that wed strong research design with rigorous statistical procedures to form the basis of objective decision making. Fortunately, the importance of using rigorous research designs and statistical procedures for decision making is no longer a cry echoed in only the halls of the academy.

Public and private leaders have come to acknowledge the role of statistics for generating reliable and valid decisions, and entire industries have blossomed through the value proposition that they can provide access to such methods (Stone & Denham, 2004). The appreciation for objective evidence in the public sector is best demonstrated in the deliberate actions of numerous federal agencies to establish systematic approaches to using evidence for decision making (Maynard, 2018). The nation's first large social experiment was conducted only fifty years ago but now countless numbers of social experiments are conducted each year ("Social Experiment"). The practice of using robust statistical procedures for decision making does not appear to be disappearing anytime soon from the landscape.

As is the case with many fields of inquiry, there are distinct empirical paradigms that scholars can select when conducting their research. These distinct paradigms can diverge but they also can overlap in many respects. The esteemed work of Kuhn (1962) suggests four principles for a paradigm. Kuhn writes a paradigm must provide guidance on what is being observed, define the kinds of questions being asked, define how questions are being asked, and provide a lens for interpreting scientific results. Within the field of applied statistics, two popular paradigms dominate the landscape. These two paradigms are known as the Bayesian approach and the frequentist approach. It is worth noting that the literature on these two statistical paradigms often uses the word framework interchangeably with the word paradigm when describing the two approaches (Rupp, Dey, & Zumbo, 2004). Although there are important philosophical distinctions between the two words, for simplicity sake and to be aligned with the peer review literature, I use the term paradigm and framework interchangeably when referring to the Bayesian and frequentist approaches throughout this dissertation.

Although both the Bayesian and frequentist frameworks share a similar history in terms of when they were developed, the frameworks themselves have experienced a very different evolution in terms of their role in academy training and their use by professional statisticians.

### **Bayesian and frequentist methods: An evolving landscape**

Scholars have documented the evolution of these paradigms in terms of their role in undergraduate and graduate training as well as the reason why the Bayesian paradigm has gained popularity in recent times. This work has revealed that historically the frequentist paradigm has virtually captured the entire landscape of

statistical training, leaving little room for student exposure to the Bayesian paradigm (Parker, 2004). Most scholars also agree that the current rise in popularity for the Bayesian paradigm is due to the advent of computationally intensive computing leading to easier estimation of Monte Carlo Markov Chains (Ravenswaaij, Cassey, & Brown, 2018; Paap, 2001). The benefits of Monte Carlo Markov Chains are explained by Ravenswaaij, Casey, & Brown (2018) as allowing researchers to estimate complicated multivariate conditional distributions (e.g., reflecting the expected impact of an intervention, conditional on covariates) that historically would have been very difficult if not impossible to produce.

Typically, one would calculate percentiles of a hypothesized normal distribution using the normal distribution equation<sup>1</sup> with a given mean and standard deviation (e.g., based on sample estimates), but by using a MCMC method an individual can calculate a large number of possible random draws from an estimated conditional distribution, and then calculate percentiles from that sample (i.e., the Monte Carlo). The Markov chain has an inherent mathematical property associated with it that assumes that each random sample draw is dependent only upon the prior draw and not any other draw in the sample chain. Hence, the benefits for MCMC's become clear especially when working with difficult multivariate conditional distributions where calculating sample statistics is not straightforward. It is important to point out that MCMC approaches, although usually associated with Bayesian methods, are not exclusive to Bayesian approaches and can also be used in a

---

<sup>1</sup> The Normal Equation is expressed as follows  $Y = \{ 1/[ \sigma * \text{sqrt}(2\pi) ] \} * e^{-(x - \mu)^2/2\sigma^2}$  ("The Normal Distribution")

frequentist framework (Hamra, MacLehose, & Richardson, 2013). In sum, although frequentist methods have dominated the educational landscape, the recent advent of MCMC computing has increased popularity in Bayesian methods.

Work by Lecoutre (2006) demonstrates that the lack of training for students in Bayesian methods is problematic beyond the fact that these students might not be able to take advantage of MCMC methods. Lecoutre argues that more and more journals and publication outlets are altering their preference for the researcher to report a Bayesian framework instead of the frequentist framework. Further work by Kaplan (2018) argues that the benefits of Bayesian methods are substantial enough to warrant federal agencies to outline requirements that specify Bayesian approaches in grant applications, develop requirements for evaluating Bayesian approaches, and include Bayesian approaches in evidence reviews. Despite the current lack of training, many graduate programs and faculty have recognized the void and are active in offering commentary on their own approaches for introducing students to Bayesian methods (Gelman, 2008). In conclusion, there is promise for a future where opportunities for instruction in Bayesian analysis meet with the expected demands from journal editors and federal agencies to use Bayesian techniques.

Thus far I have argued that Bayesian techniques have typically been sidelined as frequentist techniques have dominated the educational landscape. However, advances in computationally intensive computing have enabled researchers to benefit from MCMC approaches spurring a renewed popularity for Bayesian methods. Currently there is concern that students are inadequately prepared for Bayesian analysis, but I also argue that faculty are gradually making strides to prepare their students, and there is promise for a future where students will meet this demand. If we

accept these arguments suggesting that Bayesian methods are the “wave of the future,” we must then ask, so what really is all the hype about between Bayesian and frequentist approaches? Posited a different way, what are the main differences between these approaches and more importantly, why should we care?

### **Core differences between the Bayesian and frequentist approaches**

Jordan (2009) claims that Bayesian statisticians rely on a conditional framework for generating statistical conclusions whereas frequentist statisticians rely on unconditional frameworks. Jordan understands this notion of conditionality in terms of scientific replication. A Bayesian’s goal is to make a conditional statement about their data at that current point in time and they are less interested in replicating their results. In fact, Bayesians are more interested in updating their results as more information becomes available. As more information is learned, this information is used to generate more precise estimates and inferences. On the other hand, a frequentist is more concerned with replication and aims to derive results that can be repeated.

Jordan’s (2009) second principle focuses on the role of the content expert. Jordan argues that the content expert is more important in a Bayesian analysis than a frequentist analysis. The reason for the importance of the content expert will become more evident later in this dissertation. It is important to note that the role of the content expert is not diminished in the frequentist framework. The role of the content expert in research and decision making has long been recognized although it is worth noting that recent work has questioned the utility of content experts’ participation in meta-analyses and systematic reviews (Gotzsche & Loannidis, 2012). In conclusion, although the role of the content expert is still important in both the Bayesian and

frequentist paradigms, the content expert is likely more important in the Bayesian paradigm (Jordan, 2009).

Jordan's (2009) final principle aims to describe the world views of these respective frameworks. According to Jordan, the Bayesian understands the world as an optimist, the Bayesian wants to "make the best possible use of [a] sophisticated inferential tool" (p. 2) whereas the frequentist wishes to "protect [us] against bad decisions given that our inferential procedure is inevitably based on a simplification of reality" (p. 2). Although Jordan's demarcation between these two statistical paradigms is helpful at a broad level, it is almost certainly a simplification and professional statisticians have written many manuscripts that describe the philosophical (Wilson, 2003), mathematical (Raue, Kreutz, Theis, & Timmer, 2013; Malakoff, 1999), and pragmatic (Efron, 2012) differences between these two paradigms. Recognizing that there are clear distinctions between the Bayesian and frequentists paradigms, the next logical question we must ask is why should we care about the differences between these two approaches?

### **Why care about differences between Bayesian and frequentist statistics?**

One deeply contested issue in Bayesian versus frequentist approaches is whether these approaches lead to differences in results. As with most widely contested issues in science, the answer is, it depends on who you ask. One perspective argues that both the Bayesian and frequentist approaches will lead to different answers because they provide different kinds of information within a hypothesis testing framework. The Bayesian approach allows for information to be obtained about the alternative hypothesis whereas the frequentist approach allows only for information to

obtained about the null hypothesis<sup>2</sup> (i.e., either reject the null or fail to reject it), and given this difference, one should expect that different results will be obtained from the same data (“Bayesian vs Frequentist Approach: Same Data, Opposite Results”).

However, not all statisticians agree with this answer.

Several statisticians argue that Bayesian versus frequentist methods should not lead to a difference in results. They argue this holds even in instances where different information is provided regarding the hypothesis (i.e., the Bayesian provides information on the alternative and the frequentist provides information on the null), because simple calibration techniques can be applied to establish similar results (Silva, 2017). Even if we assume consensus can be reached that these approaches do not produce inherently different results, the question remains whether or not the interpretation and/or presentation of results and subsequent decision action by individuals reviewing evidence (e.g., policy makers, legislative aides) might differ across these paradigms. As mentioned above, one must be highly trained in statistics to recognize that these approaches provide different information about hypotheses and thus, we should not be surprised that the subtle but complicated differences in these two approaches might in turn result in different decision actions. Even professional researchers struggle sometimes with interpreting statistical concepts.

---

<sup>2</sup> An alternative framework known as the Neyman-Pearson lemma under the frequentist paradigm argues that a hypothesis can be accepted on the basis of data and therefore diverges from the traditional Fisher frequentist interpretation that one can only *fail to reject* or *reject* the null hypothesis. Since the Neyman-Pearson framework has not dominated traditional frequentist approaches like the Fisher framework, the focus of this project will be on Fisher’s frequentist framework and the Bayesian framework. For more information on the Neyman-Pearson framework (Dantzig & Wald, 1951).

Past research has shown that both professional researchers and undergraduate social science majors struggle with concepts in statistics. Work conducted by Hoekstra, Morey, Rouder, and Wagenmakers (2014) asked professional researchers and undergraduate psychology majors to interpret six statements about confidence intervals. Despite all six statements being false, their work found that on average both professional researchers and undergraduates agreed with three or more statements, which the researchers interpret as evidence for a robust misunderstanding of confidence intervals. Thus, if both professional researchers and undergraduates struggle with a statistical concept like confidence intervals, one can reasonably assume that problematic interpretation may result when individuals are asked to interpret complex nuances between the Bayesian and frequentist approaches.

I stated earlier that Silva (2017) argues that the presentation of results under a Bayesian or frequentist framework should not lead to differences in decision making. Although this supposition may be true, the claim itself is a normative one and it takes for granted underlying differences in how both professional researchers and educated individuals (e.g., undergraduates) understand and make decisions about statistical information. As shown earlier, both professional researchers and undergraduates can be subject to erroneous reasoning about statistical information. Recognizing the risk inherent for individuals interpreting statistical information, researchers have called for results to be presented in the most accessible way possible to guard against erroneous interpretation.

Scholars have argued that we should present statistical information in the most accessible manner possible (May, 2004; Epstein, Martin, & Schneider, 2006), and that Bayesian methods lead to more interpretive ease than frequentist methods (Buchinsky

& Chadha, 2017; Gurrin, Kurinczuk, & Burton, 2000). If we accept this claim, then we are faced with a natural extension of the earlier issue with respect to the interpretation of statistical information. Is it the case that Bayesian methods are indeed easier for individuals to interpret and less prone to misinterpretation?

What are the implications of a framework that is easier to interpret? Might individuals be more likely to endorse support for program or policy under a framework that is easier to interpret? If it is the case that individuals are more likely to endorse programs or policies under a Bayesian framework, can this simply be explained by the fact that they find the Bayesian results easier to understand? Maybe the optimist versus pessimist distinction suggested by Jordan (2009) is more than just a useful dichotomy for understanding Bayesian versus frequentist statistics.

Perhaps, the presentation of Bayesian results taps into an aspect of cognition known as the optimism bias. This bias has long been documented in the behavioral economics literature and numerous experiments have shown that in general humans tend to be more optimistic than pessimistic across different domains such as showing overconfidence in their ability to remain married, underestimating the likelihood of a serious car accident, and substantially underestimating their lifetime risk for cancer (Sharot, 2011). Early empirical work has only begun to explore these many questions.

### **Early empirical evidence on Bayesian decision making**

Early empirical evidence has just begun to answer the aforementioned questions. Policy researchers Chandler, Martinez, Finucane, Terziev, & Resch (2019) conducted a study with a convenience sample of Amazon Turk users. In their study, they presented individuals with statistical results from a fictitious impact evaluation of

a new educational technology under both a Bayesian and frequentist framework. In one of their experiments, they varied the presentation of these results in terms of strong, moderate, and weak evidence providing corresponding p-values (frequentist scenario) to posterior probabilities (Bayesian scenario) for each one of the evidence scenarios. Their survey items asked respondents questions about their willingness to endorse a new technology, how confident they were in their decision to endorse, and which statistical paradigm (Bayesian or frequentist) afforded more interpretive ease.

Their results were striking and suggest individuals are more likely to endorse an education technology when it is presented under a Bayesian framework for both the strong and moderate evidence scenarios despite being shown equivalent impact data under a frequentist framework. For the weak evidence scenario, there was no differential endorsement for support of the program based on statistical paradigm. Furthermore, respondents felt more confident in their decision when results were under a Bayesian framework and felt the Bayesian framework was easier to interpret. In recognition of this compelling early research, I argue that the Bayesian and frequentist paradigms do cause a difference in interpretation at least when presented to non-statisticians and seek to expand this work in meaningful and important ways.

### **The case for studying statistical judgment**

One simple reason we need a better understanding of the role Bayesian statistics might have in influencing statistical judgment is, as argued earlier, Bayesian results are proliferating many fields of inquiry. Although research into statistical judgement for high stakes decision makers (e.g., medical doctors, superintendents, and Congressional staffers) should be a priority, Congressional aides are of concern due to the amount of information they are often forced to consider and the consequences of

making an incorrect decision. As such, legislative aides are the population of interest for this dissertation project.

Individuals unfamiliar with legislative decision making often assume it is their elected officials, the members of the House and the Senate in the United States of America who cast votes representing the interests and needs of their constituents on important issues. Although this is technically true, these elected officials are often not the primary consumers of the actual evidence that is used to inform their decision. Legislative aides provide their Congressional employers with briefings that summarize findings from empirical studies. These briefings are usually distillations of complex evaluation reports but still contain findings that require statistical interpretation and ultimately, a decision recommendation. Furthermore, Members of Congress typically view conversations with their aides as one of the most influential factors for how they eventually cast their vote, and anecdotal evidence suggests many aides actually understand the issues surrounding a vote better than the Representative or Senator they work for (Tolchin, 1991). Recent innovative quantitative work studying the role of the legislative aide has provided a new level of evidence regarding their influence.

Political scientists Montgomery & Nyhan (2017) conducted a groundbreaking study using a network analytic model to quantify the role of the aide in their Member's decision process. They found that aides are highly influential on a Member's legislative effectiveness and voting pattern. Specifically, when aides leave one Member's office and join another, they influence the voting pattern of that new Member in striking ways. These researchers interpret the evidence to suggest that aides have an astonishing effect on their Member's voting pattern in ways that exceed prior thinking of political scientists.

Recognizing that legislative aides are often not professionally trained statisticians and lack the tools necessary to reason robustly about statistics is not a novel revelation for policy professionals. A recent employment survey of legislative aides found that almost 40 percent of all legislative aides have only bachelor's degrees and the top three majors for legislative aides are political science, law, and business which represent over 55 percent of the reported majors by aides ("Working As A Legislative Assistant"). These demographics suggest aides are part of a class that may be vulnerable to erroneous statistical reasoning. In conclusion, recognizing both the role aides have in the Member's decision process and their susceptibility to erroneous statistical reasoning, I argue they are a high priority population and are worth investigating whether or not statistical paradigm presentation affects their decision action.

### **High stakes decision making: More than just evidence**

Legislative aides are asked to make recommendations about a program or policy not based solely on the level of evidence provided by the impact evaluation. Factors with respect to the cost of the program as well as issues regarding implementation feasibility must always be considered alongside impact evaluation results. Thus, although prior research has shown individuals are more likely to endorse an educational technology when results are presented under a Bayesian framework, it is unclear how individuals will respond when other important factors like cost of the program and implementation feasibility are included as salient information in the decision process.

Prior research supports the claim that the cost of a program and implementation feasibility of that program are important concepts in legislative

decision making. Cost of the program is defined as the monetary resources required to implement the program and implementation feasibility is understood by how difficult the program itself is to implement. For example, a new math curriculum might be very inexpensive, but the instructional material so complicated that the teachers are unable to implement it with their students. Similarly, a program might be very expensive to purchase, but is very easy to implement from the perspective of the practitioner tasked with implementing the program. These various permutations demonstrate the complicated relationship that exists between cost and implementation feasibility of programs.

Guidelines for school districts provided by the Institute for Education Sciences exist to help districts balance the demands of evidence, cost, and implementation feasibility when deciding whether or not to endorse a new program. Work conducted by Hollands & Levin (2017) provides a useful framework for school districts to understand the role of these factors. These authors describe four frameworks: the cost-feasibility framework where an assessment is made on the basis of cost only—does the district have enough dollars to fund the program? A cost-effectiveness framework where the decision is determined based on the proportion of the cost per student over the gain in ability expected by that student. A cost-benefit framework where the cost decision is expressed as a proportion of dollars spent now versus dollars saved in the future. The final framework is a cost-utility framework where cost is understood as a function of the utility it brings to the population (e.g., it is a highly desired program by parents and teachers). The various applications of these cost analyses are beyond the scope of this dissertation but the point nonetheless is to suggest that cost and implementation feasibility are important features of the

Legislative decision process and must be understood always in the context of evidence strength.

In conclusion, one cannot divorce cost or implementation feasibility from evidence, and it is the tension between these three factors that must be resolved. Therefore, although early empirical work in this field has made a substantial contribution, I seek to expand this work in meaningful and important ways by examining not just whether Bayesian and frequentist methods lead to differences in judgment for a high stakes decision population like legislative aides, but also what role cost and implementation feasibility have in the decision process, and how that might interact with the Bayesian/frequentist effects.

In this dissertation project, I aim to answer the following three research questions (RQ's) and offer the following hypotheses (H).

RQ1: Do the Bayesian and frequentist paradigms result in differential judgement in terms of whether or not legislative aides are willing to endorse an education program? In other words, are aides more willing to endorse a program when results are presented under a Bayesian framework as opposed to a frequentist?

H1: I hypothesize legislative aides will be more likely to support funding for a program when results are presented in a Bayesian framework as opposed to a frequentist.

RQ2: What is the role of cost information for legislative aides when determining whether or not to endorse an education program presented under a Bayesian versus frequentist framework? In other words, are aides more willing to spend additional dollars when results from a program evaluation are presented under a Bayesian framework, instead of a frequentist?

H2: I hypothesize legislative aides will be more likely to support funding for more expensive education program when results from a program evaluation are presented under a Bayesian framework.

RQ3: As argued earlier, implementation feasibility is important alongside evidence and cost in policy decision making. A program that has robust evidence and is low cost but otherwise extremely difficult to implement, may be unlikely to gain support. For RQ3, I ask what is the role of implementation feasibility for legislative aides when determining whether or not to endorse an education program presented under a Bayesian versus a frequentist framework?

H3: I hypothesize legislative aides will be more likely to endorse a program that has less feasible implementation when results from the impact evaluation are presented under a Bayesian framework.

## Chapter 2

### LITERATURE REVIEW

#### **A Very Brief Tale of Two Statistical Paradigms**

The Bayesian and frequentist statistical paradigms have dominated the field of statistical inference for many years. In fact, the discovery of these paradigms can be traced to identical centuries. The esteemed mathematician Thomas Bayes is credited with the formulation of Bayes' theorem that serves as the foundation for the Bayesian statistical paradigm (James, 2019). Bayesian statisticians understand the world through increasing degrees of certainty where statistical results are presented in probability statements.

The origin of the frequentist paradigm is shared across several mathematicians. Confidence intervals were first credited to the mathematician Abraham de Moivre (Stahl & Johnson, 2007) but were forgotten about for almost a full century until the mathematician Pierre-Simon Laplace revisited the role confidence intervals have in statistics. Laplace is also credited with the discovery of p-values (Wilcox & Serang, 2017). However, it is Fisher (1954) who is credited with the popularization of p-values within the frequentist paradigm and their use in contemporary statistics. Later on I will provide a more thorough discussion of these statistical concepts, but before I begin, I will examine the role of hypothesis testing in statistical inference.

#### **Hypothesis Testing for Inference in Experiments**

Scientific investigation begins with the development of a hypotheses. The literature that discusses how science defines a hypothesis is beyond the scope of this dissertation but worth mentioning is that both applied social scientists and philosophers of science disagree considerably on the very definition of a hypothesis

(Pigliucci, 2009). For the purposes of this dissertation, I use the commonly accepted definition by the American Psychological Association that states that a hypothesis is “an empirically testable proposition about some fact, behavior, relationship, or the like, usually based on theory, that states an expected outcome resulting from specific conditions or assumptions” (American Psychological Association [APA], “hypothesis”). Since there is substantial disagreement on what counts as a hypothesis, there should be no surprise that there is also substantial disagreement on what counts as an experiment (Andersen & Hepburn, 2015). I also use the American Psychological Association’s commonly accepted definition of an experiment that defines an experiment as:

“a series of observations conducted under controlled conditions to study a relationship with the purpose of drawing causal inferences about that relationship. An experiment involves the manipulation of an independent variable, the measurement of a dependent variable, and the exposure of various participants to one or more of the conditions being studied. Random selection of participants and their random assignment to conditions also are necessary in experiments” (American Psychological Association [APA], “experiment”).

These arguments over defining an experiment are again beyond the scope of this dissertation but acknowledging the definition provided by the American Psychological Association enables common ground for a further discussion on the role of hypothesis testing in social science experiments.

Hypothesis testing for experiments involves generating two distinct ideas about what might happen with the data. These two types are referred to as the null hypothesis (H<sub>0</sub>) and the alternative hypothesis (H<sub>1</sub>). The null hypothesis states that

there is no association between two measured phenomena, and the alternative hypothesis states there is an association between the two measured phenomena (Fisher, 1954). Examining this under an experimental framework, a Bayesian would present the likelihood of a null hypothesis as a probability statement given the distribution of the experimental results. Bayesians are allowed to incorporate prior probabilities into their computation when specifying hypotheses, which serves as a key difference between the Bayesian and frequentist paradigms. Bayesians are also allowed to provide evidence in support, or not in support of, the alternative hypotheses whereas such inferential reasoning is not allowed in the frequentist tradition.

The null hypothesis statistical testing (NHST) framework is the dominant framework used in the frequentist tradition and allows only for inference regarding null hypotheses. Frequentists may only reject or fail to reject the null hypotheses, and frequentists are not allowed to state an inference regarding the alternative hypothesis, although this limitation of the NHST framework is commonly ignored or not well understood (Gliner, Morgan, Leech, & Harmon, 2001). It is important to note that Bayesian hypothesis testing is useful in many situations especially when researchers are interested in providing evidence in support of the alternative hypothesis,  $H_1$ . The Bayesian framework becomes potentially less useful when researchers are more interested in providing evidence regarding the null hypothesis,  $H_0$ .

### **Concepts Relevant to Decision Making for Statistical Inference**

#### **P-Values, Power, and Posterior Probabilities**

In the frequentist paradigm, a p-value is basically a decision aid for determining how to act. A commonly accepted p-value is 0.05. Frequentists reject the

null hypothesis if it is  $< 0.05$  and fail to reject the null hypothesis if the outcome is  $> 0.05$ . As discussed earlier, frequentists never accept or reject their alternative hypothesis, they always either reject or fail to reject the null hypothesis. Assuming the null hypothesis were true, the p-value provides “the probability of obtaining a result at least as extreme as the one that was actually observed” (Panagiotakos 2008).

In the frequentist paradigm, we never really know the true value of the parameter. Instead, we estimate the value of the parameter and thus, the frequentist paradigm always produces an error rate where we would otherwise fail to reject the null hypothesis. P-values are a reflection of type I error rates that is the probability of producing a false positive. A false positive occurs when we reject the null hypothesis but should have failed to reject the null hypothesis. The type I error rate is the alpha value (typically set to .05). In recent times, p-values have come under heavy criticism from professional statisticians and it is now accepted that p-values should never be used solely to inform science, business, or policy related decisions (American Statistical Association [ASA], “Statement on Statistical Significance and P-Values”). Power is another important concept in the frequentist framework.

Power is the chance of correctly rejecting the null hypothesis when the null hypothesis should be rejected. Power is calculated by subtracting the beta value from a base of 1 (i.e.,  $1 - \beta$ ). Similar to the earlier described alpha value, beta is an expression of type II error, which is the probability of producing a false negative. A false negative occurs when we fail to reject the null hypothesis but should have rejected the null hypothesis. Beta is typically set apriori to 0.20 and thus the commonly accepted standard for statistical power is 0.80. In the frequentist tradition, there are tradeoffs between varying levels of alpha and beta.

Researchers typically compute sample size requirements based on different levels of assumed alpha and beta as well as assumed levels of a minimum detectable effect size. A more substantial discussion on the concept of effect size is presented later. These computations regarding sample size provide researchers with an estimate of the number of participants needed for data collection in order to achieve the previously calculated alpha and beta levels, or conversely, precision and power. Depending on the research program and the availability of resources, different levels for alpha and beta are necessary. For example, a NASA engineer may set an alpha level to 0.00001 because even a small increase for type I error might have catastrophic consequences. Alternatively, in a psychology experiment, a researcher may be willing to accept greater type I error and increase the alpha level to 0.10.

Beta has a similar tradeoff where some research programs mandate higher levels of power (i.e., 0.90, beta = 0.10) whereas others are willing to accept lower levels (i.e., 0.70, beta = 0.30). Other reasons why alpha and beta levels might change is due to cost to collect the data. Costs incurred with data collection are a common reason a researcher might adjust alpha and beta levels and grant applications sometimes request a discussion of these tradeoffs. Although these concepts of the p-value and power are essential to a frequentist framework, these concepts do not exist in a Bayesian framework. Bayesians focus on the relative influence of the prior and the data likelihood on the posterior probability rather than p-values and power. In other words, Bayesians understand the concept of precision through the selection of the most robust available prior.

Although the concept of a posterior probability might seem abstract, the mathematical expression of the Bayesian is actually simplistic:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

As mentioned earlier the development of this equation was attributed to Thomas Bayes. This equation states that the probability of event ‘A’ is dependent upon the known or assumed probability of event ‘B’. Orloff & Bloom (2014) present a helpful conceptualization of the Bayes’ formula for modern statistical hypothesis testing:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

In the above formula,  $H$  is the hypothesis we seek to test and  $D$  represents the data that will be collected to determine whether or not there is support for the hypothesis.  $P(H)$  is meant to represent the *prior*, which represents the prior probability of the hypothesis being true before any data is collected.  $P(H|D)$  is known as the *posterior*, which represents the probability of the hypothesis being true after the data is analyzed.  $P(D|H)$  represents the likelihood of observed data given the prior  $H$ , and  $P(D)$  is the complete probability of the data after consideration of all the hypotheses. The outcome of this equation is known as the *posterior probability* and this outcome provides the information necessary to either reject or accept the null hypothesis.

If researchers were conducting an experiment where all possible prior probability distributions were completely known (e.g., a coin flip experiment) then the posterior probability of the distribution would be predicted with 100 percent accuracy. This allows Bayesian analysis to cross the inductive-deductive gap and although an interesting theoretical concept, in applied social experiments, the prior probability is

never known perfectly and often not known at all, which leads to one of the main criticisms of the Bayesian approach. Namely, it requires specification of a parameter's prior distribution. Bayesians have two types of priors known as objective priors and subjective priors.

Bayesians make distinctions between subjective priors (i.e., priors that are selected in lieu of empirical data) and objective priors (i.e., priors that are selected on the basis of empirical data). The selection of priors is one of the most prominent targets of criticism by frequentist loyalists. Subjective priors are argued by some to be nonsensical and are heavily criticized as opening the door to bias (Gelman, 2008); however, organizations are putting forth guidelines to help standardize the calculation of subjective priors (Vose, 2017, "Subjective Priors"). Further, I mentioned earlier in this dissertation that the role of the subject matter expert is more important in the Bayesian framework than the frequentist, and this is especially true with subjective priors. The subject matter expert can help inform the robustness of the subjective prior. Objective priors are largely considered better than subjective priors since they are drawn from empirical data, but they have also been criticized since the quality of the prior data determines the quality of the objective prior.

One intuitive way to understand the prior is through the example of medical diagnostic testing. Medical diagnostic tests are always associated with a sensitivity score (i.e., the ability of the test to correctly identify those with the condition) and a specificity score (i.e., the ability of the test to correctly identify those without the condition). No medical diagnostic test has perfect sensitivity and specificity. When assessing the likelihood that a given member of the population has a specific condition, one must always consider the base rate (i.e., the likelihood that an

individual has a condition absent any other mediating factors). When calculating the likelihood of an individual having a specific condition, the base rate must be considered. If the base rate is ignored, this is considered the base rate fallacy and individuals who make this fallacy will not accurately report the likelihood of an individual having a certain condition. The Bayesian prior works in a similar manner and can be understood as representing the base rate.

This distinction of the prior is the hallmark difference between a Bayesian and a frequentist with respect to hypothesis testing. A frequentist would always state that prior information about the certainty of a hypothesis is unknown, as a frequentist believes that is the purpose of collecting data to test the hypothesis. A Bayesian would dismiss that claim as being ridiculous noting that we all have prior expectations for how phenomena behave, and we should incorporate that information into our hypotheses.

### **Concepts Relevant to Precision for Statistical Inference**

#### Standard Errors, Confidence Intervals, and Credible Intervals

Confidence intervals and effect sizes are now considered equally important in the frequentist framework as is the concept of the p-value. In fact, many scholars argue that researchers should report confidence intervals and effect sizes over p-values (Nakagawa & Cuthill, 2007) and some journals even mandate this practice (Trafimow, 2018). As such, a thorough understanding of these concepts in the frequentist paradigm is important. Unfortunately, confidence intervals have robust misunderstanding even by professional researchers (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Given a confidence interval of 0.6 to 0.82, a common

misunderstanding of confidence intervals can be expressed as follows, there is a 95% probability that the true value of the population's parameter falls within that confidence interval. This statement although intuitive, is indeed false. A confidence interval provides an estimate of the procedure used, not of the result. An appropriate articulation of a confidence interval is that 95% of the time, repeated samples (assuming a similar sample with similar sampling procedures) will produce confidence intervals that contain the true population mean. However, since no two samples are identical and since each time we draw a random sample, what constitutes that sample changes, the resulting interval will change. A study that is conducted 100 times will produce 100 different confidence intervals, for which 95 can be expected to contain the true value of the population parameter's mean.

The confidence interval contains both a lower bound and upper bound estimate. Frequentists derive these lower bound and upper bound estimates by multiplying the standard error of the parameter's estimate by  $\pm 1.96$ . The number 1.96 is derived from the assumptions associated with a normal distribution. Under a normal distribution, 95% of the data will lie within 1.96 standard deviations of the mean. A standard error is a measure of variability in a statistic and is computed by dividing the standard deviation for the sample by the square root of the sample size. Multiplying the standard error by  $\pm 1.96$  establishes a lower and upper bound range for where the parameter's true mean value is estimated to fall assuming a similar population and when similar sampling methods are used. The confidence interval is thought to provide much needed context for a p-value. Unlike with p-values and power where these concepts do not even exist in a Bayesian framework, Bayesians have a similar concept to confidence intervals called credible intervals.

Bayesian estimates produce probability statements for a parameter and a credible interval provides much needed context with respect to a range of values for the probability's result. Similar to the 95% confidence interval used in the frequentist framework, a 95% credible interval is also commonly assumed in the Bayesian framework. Assuming a credible level of 95% was set and an interval of .75 to .80 was observed, a Bayesian would express the credible interval that there is a 95% chance the population's true mean value lies between .75 and .80. The mathematical expression of the credible interval is expressed as  $(P(a \leq \theta \leq b|x = 1 - \alpha))$ . It is important to note the difference in interpretation here; the Bayesian makes a statement about the true value of the population mean within this fixed interval. For the Bayesian, the parameter's value might vary but will do so within the interval. In the frequentist approach, the parameter's value will not vary, it is the interval that will vary but the true parameter value can never be known with certainty.

Although credible intervals may seem to share a common concept with confidence intervals, it is important to highlight the stark contrast described above. Confidence intervals treat the estimate for the parameter as fixed and the confidence bounds as random, but Bayesian credible intervals treat the estimate for the parameter as fluctuating and the interval as fixed. This is an important philosophical difference and worth noting as this difference is what leads to the previously described distinction in interpretation.

As mentioned earlier, effect sizes are now considered equally as important as p-values for decision making. Recently scholars have advocated for researchers to conduct tests of equivalence within the frequentist framework for making inferential decisions in addition to considering the value of effect sizes and confidence intervals

(Lakens, 2017). Thus, a discussion of these additional concepts is necessary alongside with the analogous Bayesian concept, regions of practical interest.

### **Additional Concepts Relevant to Decision Making for Statistical Inference**

#### Effect Sizes, Tests of Equivalence, and Regions of Practical Interest

Effect sizes are also useful to provide context to p-values. Examples of effect sizes are the  $R^2$  statistic or the popular standardized Cohen's  $d$  that is calculated as the mean difference between two groups divided by the combined standard deviation ("The Cohen's  $d$  Formula"). Effect sizes can also come from unstandardized mean difference comparisons or regression coefficients. Effect sizes provide important information for decision makers when deciding on how to act upon the results of an experiment.

Misunderstandings aside, effect sizes nonetheless can be used to provide helpful information regarding whether or not to fund a policy or program, especially given a p-value near the significance threshold. Take for example results from a recent education study that sought to improve children's mathematics scores. A p-value that is trending towards significance was obtained,  $p = 0.08$  but a large Cohen's  $d$  value of 0.86 was computed. A strict traditionalist interpretation under a frequentist framework would indicate failure to reject the null hypothesis and in turn, might warrant the decision maker not to support funding for the program. However, the large observed effect size provides much needed context to the p-value and some scholars suggest we should consider support for programs where p-values trend towards significance but large effect sizes are obtained. In other words, the sizable benefit of the program is worth the small increased risk of a type I error. The

Bayesian framework also includes the concept of effect sizes and again, these effect sizes are expressed with a resulting probability. A similar formula is used to calculate the mean effect size for a Bayes factor,  $(\mu - 0 / \sigma)$  and a resulting probability is placed around this observed effect size.

As discussed earlier, frequentist decision making requires several pieces of information, namely p-values, power levels, confidence intervals, and effect sizes. Typically, these are the pieces of information used to make decisions in a frequentist framework, but more recent times have also called for the use of tests of equivalence. Equivalence tests allow researchers to provide evidence for the absence of a meaningful effect by conducting a test of equivalence. One such equivalence test is a test of one-sided equivalence (TOST). In the TOST procedure, an upper and lower bound equivalence is specified based on the smallest potential effect size of interest to the researcher (e.g., .20 standard deviations). If both the lower bound and upper bound test can be rejected the researcher can then say the smallest effect size of practical interest can be rejected. TOST procedures are specifically useful for p-values trending towards significance as they add an extra decision aid for the researcher. The concept of the TOST in the frequentist framework is similar to the concept of regions of practical interest (ROPE) in the Bayesian framework.

Regions of practical interest (ROPE) is much more pronounced in the Bayesian framework and commonly reported for all studies published using Bayesian methods than the comparative TOST is within frequentist method. The ROPE provides an added decision tool for understanding if an observed posterior distribution and its associated credible interval is of *practical interest*. Similar to the TOST, ROPEs state a range for where the value of the outcome produces meaningful practical interest.

Take for example an education intervention where the students are expected to increase their learning by 2 to 10 points. Using a Bayesian ROPE, the researcher would set their region of practical interest between this range of 2 to 10 points and compute a high-density interval (HDI) of 95%. The HDI creates an interval for which the majority of the distribution can be understood as falling within that interval. Kruschke (2015) suggests the HDI is the most credible way of expressing the interval because it covers the greatest amount of the distribution and as such, any part of the distribution within the HDI can be thought to have higher credibility than any point not within the distribution. If the value for the ROPE falls within the HDI, then Bayesians can accept that the education intervention is of practical interest whereas if the value of the ROPE falls outside the HDI, Bayesians can reject the education intervention as not being of practical interest. It is important to note that this method is not as straightforward as I just described for decision making, because Bayesians disagree about where to set their ROPE values and also disagree about what number should define the HDI. Some Bayesians argue an HDI of 90% or 96% should be used instead of 95%. ROPEs are nonetheless helpful as part of the decision process.

### **Revisiting Core Differences: Bayesian v. Frequentist**

Earlier in this dissertation I described three core theoretical differences between the Bayesian and frequentist frameworks. Now I turn my attention to three core applied differences between these two frameworks. I argue for these three core differences because they hold the greatest implications for applied researchers deciding which framework to utilize.

Recognizing the differences between a Bayesian approach and a frequentist approach is important for understanding why one procedure might be preferred over

the other and then in turn, to understand why individuals might reason in different ways when presented with information under these two paradigms. Lindley (1993) provides an illustration for some of the benefits supplied by the Bayesian framework through the now famous tea experiment. The tea experiment involves an exchange between R.A. Fisher and Muriel Bristol. Below I present both the experiment and its implications for frequentist and Bayesian experimental analysis.

In this thought experiment, R.A. Fisher is meeting his colleague Muriel Bristol for a cup of tea. Fisher prepares a cup of tea for Bristol where he adds milk to her tea after already pouring tea from the pot into her cup. Muriel is annoyed by this practice and addresses Fisher by saying she prefers the milk added to her cup before the tea is poured into the cup. She goes on to claim that she has the ability to actually distinguish whether milk is added before or after the tea is poured into the cup. Fisher acknowledges the criticism and challenges Muriel to an experiment where he presents her with six cups of tea to determine if she indeed does possess the ability to distinguish between when the milk is added to the tea.

Fisher makes two basic assumptions when designing the analytic methods for his experiment: 1) the goal of experimental design is to discredit the null hypothesis and 2) the null hypothesis is defined as the absent of the event. In the tea experiment, the null hypothesis is defined as Muriel's inability to distinguish between cups of tea. Fisher amends his argument with two points: a) either the null hypothesis is true and the probability of events as, or more, extreme than that observed is small, or b) the null hypothesis is false. Thus, a 5% significance result "refers to the probability of all results as, or more, extreme than that observed". In the case of the tea experiment, there are seven cases (six where Muriel gets five correct and one wrong and one where

Muriel gets them all correct). Since the  $n$  trials are fixed and the outcome options are correct or incorrect, our discrete probability distribution is known as the binomial distribution. This result is formulated as  $7(1/2)^6 = .109$ . With a result of .109, we fail to reject the null hypothesis.

What if we change the above scenario to one where Muriel stops only when she gets a cup of tea wrong? In other words, instead of Muriel being presented with six cups of tea and having to make a guess at each cup, this time Muriel guesses until she gets one wrong. Through this slight adjustment, we have changed our experimental distribution to a negative binomial distribution since the sequence of the trials are not fixed to set a number but rather, the sequence continues until Muriel makes an error. In this example, Muriel correctly guesses the first seven cups of tea and incorrectly guesses the eighth cup of tea. Here the probability of the observed result is expressed as  $(1/2)^6 + (1/2)^7 + (1/2)^8 + \dots = (1/2)^6 / (1 - 1/2) = (1/2)^5 = 0.031$ . In this instance, we reject the null hypothesis that Muriel does not have the ability to detect differences in the inclusion of milk to tea.

One could argue that there is no problem with the procedure described above. The frequentist framework rests on defining the number of trials or the number of errors upfront. We either let the trials run sequentially (i.e., we do not stop until Muriel gets one wrong) or we restrict the number of trials (i.e., Muriel runs six trials); however, Lindley asserts this approach is unreasonable. Imagine if we intended to run the sequential trial but because Muriel got a call that her child needed to be picked up early from football practice because their mouthguard sensor broke, we had to cut our trials short. Why should we be concerned with a potential change to our p-value as a result of these practical concerns? Further take the most ideal scenario where we

allow chance to determine a sequential versus a fixed trial design. Suppose one were to use a fair coin to determine which experimental trial to run where heads represent the restricted trial and tails is for the sequential. The average of the two experiments is  $0.109 + 0.031/2 = 0.070$  which is not statistically significant. Suppose the coin does indeed land on tails, Lindley asks “should we really quote 0.070 merely because the coin might have showed heads?” (p. 22). Lindley believes this is absurd but absent our ability to define more extreme outcomes in Fisher’s frequentist framework, this is what we are left with unless we turn to a different framework that allows us to specify prior probabilities.

Naturally, the Bayesian framework allows us to specify a prior probability. It allows us to determine whether or not we think Muriel was guessing (i.e., a chance of 50% correct) or maybe we think Muriel has some expert powers and we want to assign a probability of 60% or even greater. The Bayesian approach has this flexibility but the frequentist approach does not.

Researchers Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2017) present an eight-step program for how traditional frequentist statisticians can become Bayesians in their article “How to become a Bayesian in eight easy steps”. The article presents several examples of differences between the paradigms and as mentioned at the start of this section, below I discuss the three most prominent differences relevant to social scientists. The first core difference is *stopping rules*, which are of most relevance to our earlier discussion on tea.

Stopping rules are understood as procedures for data collection (e.g., how the data will be collected needs to be specified in advance). Stopping data collection early under a frequentist framework presents a risk for the research project. As described

earlier, frequentist statisticians pre-specify both a type I error rate, typically (0.05) and a power level, typically (0.80). These a priori calculations include a minimum detectable effect size assumption, which in turn provides a result for the necessary sample size needed for data collection. In other words, the potential to observe the minimum detectable effect given the assumptions for type I error and power, depend upon achieving the appropriate quantity of sample. If data collection is stopped earlier, and not enough data collected, frequentist statisticians' risk not being able to observe the minimum detectable effect or having to accept higher type I error rates and reduced power levels. On the other hand, collecting and analyzing more data than necessary is also not viewed favorably in the frequentist tradition. As more data is collected, the chance of a shrinking p-value increases thus improving the chance the researcher's result will lead to rejection of the null hypothesis. Under the Bayesian framework, stopping data collection early or collecting more data than was pre-specified is not a concern from a statistical perspective since the Bayesian framework abides by the likelihood principle (that is described in the next section).

*Planned versus post hoc comparisons* are another differentiating factor.

Lindley (1993) described a core concept in statistics, the likelihood function. This function is understood as “the probability of correct classification, for the observed result” (p. 23). Denies (2011) argues that in the Bayesian framework, temporal information does not enter into the likelihood function like it does under a frequentist framework. Frequentist analysis requires the pre-specification of hypotheses for given statistical comparisons that are called planned comparisons. Anything outside of these types of comparisons are considered post hoc. The reason for this distinction in the frequentist framework is because the temporal location of the data is considered an

essential component of the likelihood function. Under a Bayesian framework, this is not the case and thus Bayesians break away from traditional frequentist thinking with respect to the temporal location of data. In a frequentist model, a theory is tested, data is collected, and then the data may support the theory or warrant its revision. In the Bayesian world, the researcher is allowed to start with data to generate the theory and the relationship between data and theory exist in a simultaneous and continual relationship.

*Multiple testing* is also not an issue for the Bayesian either like it is for the frequentist. In this dissertation, I will analyze my data through an ANOVA model. By using an ANOVA model, I will need to adjust for multiple comparisons in my ANOVA procedure, because as discussed earlier, the likelihood of observing a significant p-value increases when more hypotheses are tested. Bayesians do not need to compute such adjustments. As described earlier, a Bayesian computes a posterior probability distribution for the phenomena being tested, which is based on a prior distribution and the data likelihood. Bayesians concern themselves with computing appropriate priors rather than adjustments for multiple hypotheses because Bayesians do not have the concept of a p-value. Instead of worrying about false positive and false negatives and making adjustments to p-values and significance levels, a Bayesian analysis incorporates all the information into the posterior (e.g., it reflects the fully conditional probability distribution, including information about false positives/negatives).

Earlier I presented evidence that professional statisticians believe Bayesian statistics are easier to understand than frequentist statistics and supported this claim with novel empirical evidence that suggests college graduates agree with this assertion

that posterior probabilities are easier to understand than p-values. I also asserted earlier that the logical extension of this claim is to ask, why is this case? Now I turn to answering this question from a cognitive-developmental perspective.

### **Infants (and Adults) as Natural Bayesians**

Some scholars argue that all of human learning essentially mimics a Bayesian model (Gopnik & Wellman, 2012). Years of research in developmental psychology have uncovered that infants explore their environment by generating hypotheses and then updating future hypotheses with the information learned from their prior explorations. This framework aligns well with the Bayesian model where prior information is used to understand the likelihood of an event and then newly collected information is incorporated to update further hypotheses. Gopnik and Bonawitz (2015) present this idea of infants as Bayesians through the ‘chicken-and-egg’ problem. The age-old question of what comes first, the chicken, or the egg, has belabored philosophical discussion and public amusement for centuries. Bayesians answer this conundrum in simplistic terms—it doesn’t matter what came first. For the Bayesian, infants experience the world in the cyclic nature by which the earlier described theory to data and data to theory operates. Infants form beliefs from experience and they predict future experience based on beliefs. Adult humans experience the world in a similar manner.

Adults have built schemas that have been generated from prior experience and serve to influence their beliefs and ultimate decisions about how future events will unfold. When adults engage in tasks, these schemas are activated and provide some expectation for how the future event will unfold (Pankin, 2013). The expectations adults use to inform their future experience is informed by a prior that may be well

known, just a wild guess, or biased based on prejudice. Once the experience is complete, adults update their prior mental model that in turn serves to influence how future events that are similar, yet still novel experiences are understood. Bayesian statistics operate in this exact manner. Prior information is incorporated into a statistical model that when new empirical information is obtained, the model is updated via a posterior that in turn, can later be used as a prior distribution in future testing.

### **A Final Word on Training to be Bayesians**

Despite the long-standing tradition of both paradigms, the frequentist paradigm has for many years been the dominant paradigm in statistical training for social scientists. Up until the fall of 2018, no course at the University of Delaware, was offered in applied Bayesian analysis for doctoral students in Education. Courses that covered the mathematical properties of Bayesian methods were available in the mathematics and statistics departments, but the point here is that no course was offered on applied Bayesian methods. As mentioned earlier, the reason why the frequentist paradigm has been dominant in statistical training is a widely debated topic and beyond the scope of this dissertation. Many credit the accessibility of Fisher's work as being the reason why the frequentist tradition dominated for so many years, combined with the lack of computationally powerful machines to run Bayesian models. Nonetheless, this is an unsettled area of scholarly inquiry where new insights are certain to enter the landscape within the next several years.

Despite the long-standing history of the frequentist paradigm, as I described earlier, recent years have seen an uptick in the use of Bayesian methods for statistical inference in the social sciences. Many scholars have speculated on the reason why

Bayesian methods are growing in popularity with some arguing that the ease of Bayesian interpretation is the reason for its increased popularity, and others arguing that it is simply computational efficiency that has led to the rise in popularity. The underlying reasons aside, the reality remains that the use of Bayesian methods for generating statistical inference is gaining prominence across the social sciences.

Public policy research, defined as research using rigorous methodology from the social sciences aimed at generating evidence on the effectiveness of public programs and policies, is relying on Bayesian methods more often to analyze and disseminate results. In such a high stakes field where the results of an impact evaluation inform decisions such as whether or not to fund programs that provide services to impoverished mothers, or programs that provide job training to disabled individuals, it is important to understand the role frequentist and Bayesian statistics might have in influencing the decision makers who must make judgements based on statistical information. The claim that probability statements are easier to understand than confidence intervals and p-values is one that can be tested empirically. As discussed previously, early empirical work has already started to investigate this very hypothesis in the context of whether or not individuals differ on their endorsement of a new education technology when results are presented in a Bayesian versus frequentist framework. This early work provided empirical evidence that individuals are more likely to endorse the technology when results are presented under the Bayesian framework.

In conclusion, Bayesian methods and frequentist methods were developed approximately at the same time in history; however, during the 20th century when graduate education became accessible in the Western part of the world, Bayesian

statistics took a back seat to their frequentist counterparts. But now, during the second decade of the 21<sup>st</sup> century, a very recent shift in this trend emerged with Bayesian statistics gaining popularity and being used in the presentation of results to a wide audience of decision makers. Accepting this trend opens the door to an interesting and important scholarly question regarding how individuals make statistical judgements when presented with results under a Bayesian framework versus a frequentist framework. Early scholarly work in this field suggests that (a) individuals are more likely to endorse a technology when results are provided under a Bayesian framework, (b) they feel more confident in that endorsement, and (c) they also perceive the Bayesian presentation as easier to understand (Chandler, Martinez, Finucane, Terziev, & Resch, 2019). Recognizing the potential for Bayesian methods in the future, additional research investigating the role the framework has for influencing the judgement of actual high stakes decision makers is an important and logical next step for scholarly research. Before a discussion of this dissertation's methodology can proceed, I present a conceptual model developed for this dissertation project.

### **Statistical Judgement Model (SJM): Bayesian versus Frequentist**

For this project, I developed a conceptual model to understand the role Bayesian versus frequentist statistics have in influencing the ways in which individuals make decisions provided statistical information. The model is presented as figure 1, and I title this model the statistical judgment model (SJM). This model seeks to explain how adult humans make decisions when presented with statistical information under two popular statistical frameworks.

In the model, statistical information from a program or policy evaluation with equivalent evidence strength, defined as a medium effect size, with equivalent p-value

(0.25) to posterior probability (87.1%) is presented<sup>3</sup>. This evidence information is then supplemented with information about the cost of the program or policy. In other words, the successful endorsement of a program or policy is gated by the level of cost for that particular program or policy. The result is four possible conditions: 1) a high cost, easy to implement condition, 2) a high cost, difficult to implement condition, 3) a low cost, easy to implement condition, and 4) a low cost, difficult to implement condition. Since each condition can be presented in frequentist or Bayesian terms, the result is eight potential conditions. Due to statistical power concerns elaborated later in the methods section, I aim to test only four conditions. These four conditions are separated into two distinct study arms and presented in figure 2.

Study arm A is for high cost programs and policies whereas study arm B is for low cost programs and policies. The paths then diverge based on the level of implementation feasibility. Study arm A describes programs where implementation feasibility is easy and study arm B describes programs where implementation feasibility is difficult. Both study arms then diverge based on statistical paradigm (i.e., Bayesian versus frequentist). The result is four experimental conditions.

I hypothesize that when statistical information is presented under study arm A, high cost programs or policies that are easy to implement, the Bayesian presentation will lead to moderate support and the frequentist presentation will lead to little or no support. The underlying reason for this assumption is because I propose Bayesian presentations tap into the earlier described optimism bias and as such, individuals will

---

<sup>3</sup> For more information on the calculation of p-value to posterior probability see Chandler, Martinez, Finucane, Terziev, & Resch, 2019

be more likely to lend support even though high cost usually produces dissuasion from a program being implemented.

I hypothesize that when statistical information is presented under study arm B, low cost programs or policies that are difficult to implement, the Bayesian presentation will lead to a strong endorsement for the program or policy, but the frequentist presentation will lead to a moderate endorsement. As argued for with the study arm A scenario, Bayesian results tap into our optimism bias and thus, I believe individuals will lend strong support when faced with a program that is low cost but difficult to implement if results are presented in Bayesian terms. For programs meeting this criterion in the frequentist paradigm, I argue they will receive moderate support because individuals are more likely to implement a difficult program if the cost of that program is minimum.

Figure 1 Statistical Judgment Model: Bayesian v. Frequentist

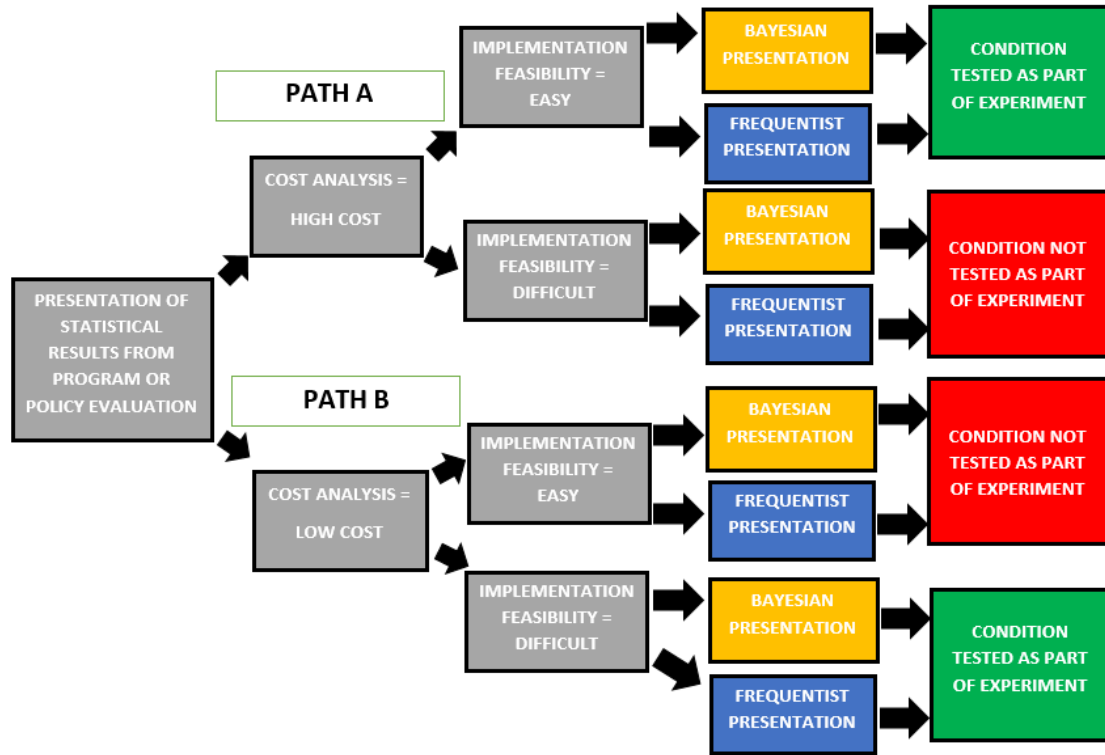
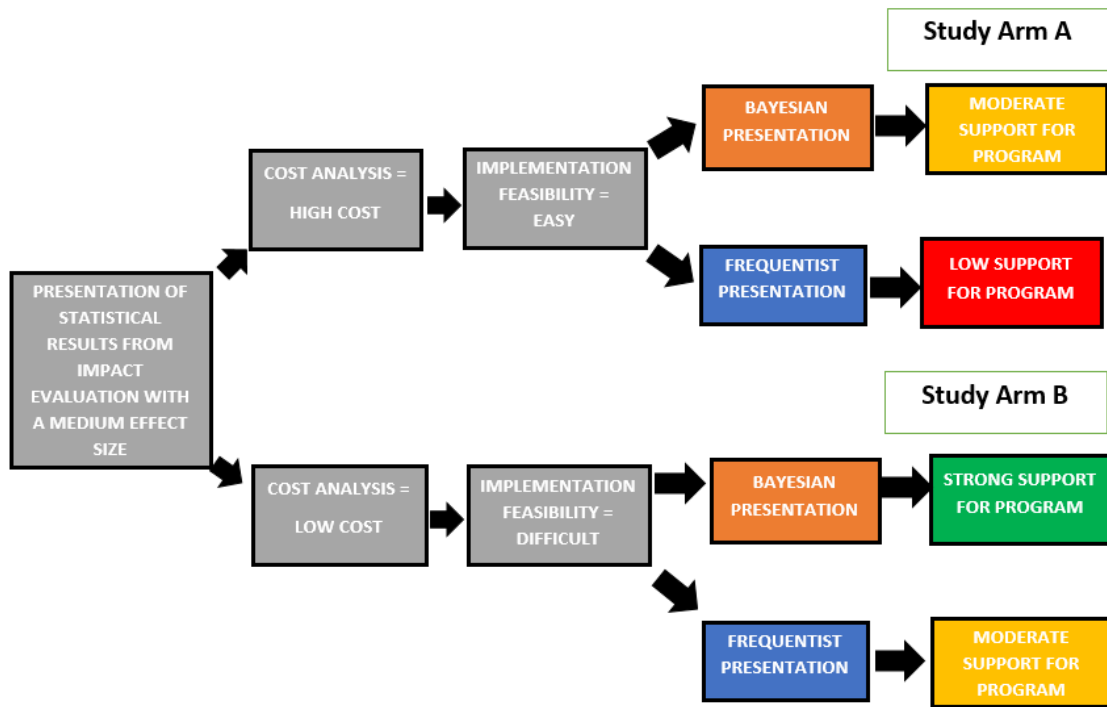


Figure 2 Presentation of Study Hypotheses by Arms Tested in this Experiment



## Chapter 3

### METHODS

#### Overview

In this section I describe the design and implementation of the survey vignette experiment employed to generate evidence for the aforementioned research questions and hypotheses. The following section outlines my dissertation methodology including the study population, recruitment strategy, the data collection tool, and the data analytic methods.

**Proposed Population.** Data for this study were collected from two target populations. The primary target population included legislative aides currently serving members of the United States House and Senate. A volunteer sample of aides were recruited from a sampling frame where information on Legislative aides was obtained from online publicly available professional networks. Utilizing professional network databases, I compiled a list of email contacts for the legislative aides. To the extent possible, information pertaining to the staffer's name, the state and congressperson who employs the staffer, the staffer email address, and the policy area for which the staffer is responsible were also collected. If the staffer's policy area was not available from the professional network, it was obtained at the end of survey questionnaire.

As an alternative to the recruitment of legislative aides, I recruited and collected data from a second population consisting of undergraduate students studying political science and international relations at the University of Delaware. At the University of Delaware, political science and international relations is one joint major. In other words, it is not possible to study one discipline without the other. Henceforth in this dissertation, I refer to subjects in this study simply as political science majors.

Although I intended to focus primarily on legislative aides, a sample of undergraduates were recruited to ensure the minimum sample size needed for statistical power (see power calculation later) was achieved. I selected undergraduates majoring in political science because these students are the ones most likely to represent actual legislative aides (“Working as a Legislative Assistant”, n.d.). In order to increase the likelihood that I achieved a sample of students who selected political science as their major, I communicated with faculty throughout the political science department at Delaware to ensure the students enrolled in the courses I visited for recruitment were likely to have already committed to major in political science. As discussed later, the recruitment speech I delivered to each class made it clear students who participated in the study must be majors in political science.

**Recruitment Strategy.** Recruitment of subjects for the legislative aide sample was conducted via email outreach after having received approval of the study protocol from the University of Delaware Institutional Review Board (IRB). Appendix A presents the IRB approved email sent to study participants asking them to participate in a short survey that examines how legislative decision makers reason about statistical information. Recruitment occurred on a rolling basis and individuals were contacted up to six times. A sample of 1,384 legislative staff was compiled and emailed inviting them to participate in the survey. A total of 30 congressional staff (2 percent) completed the survey. The University’s IRB approved the use of incentives for respondents who completed the survey.

Decades of survey research have demonstrated that incentives typically improve response rates and the use of prepaid incentives in mail and telephone surveys can improve response rates even more than post pay incentives (Singer & Ye, 2013).

However, recent research has revealed that the use of prepay incentives in web surveys does not appear to have the same benefit as mail or telephone surveys (Marken & Auter, 2017). As such, a \$10 Amazon gift card was provided to all respondents in both sample pools who completed the survey. Respondents from both sample pools were also entered into a raffle to receive a \$35 Amazon gift card. One respondent at random was selected from each sample to receive the \$35 gift card. Funds for the gift cards were provided at the personal expense of the author. The data collection period lasted six months.

The recruitment of subjects for the political science undergraduate sample occurred initially via email outreach. After receiving the appropriate approval from the University of Delaware IRB, I contacted 11 faculty in the political science department and requested a time to attend their class sessions to share my recruitment speech with students. Targeted classes were selected based on those that had the greatest likelihood of having students who already committed to selecting political science as their major. I received permission from four out of the 11 faculty to visit their classes, which equated to a total of five different class sessions. Appendix B documents the recruitment email I shared with instructors to receive their approval to recruit at their class sessions and Appendix C, documents the recruitment speech I delivered at the beginning of the political science class sessions after permission was obtained from the instructor. During recruitment, political science students who expressed interest in the survey provided their email that was then uploaded to Qualtrics and, similar to the legislative aide sample, political science undergraduates received a secure email link to complete the survey. A total of 88 students received

the survey link and 36 students (41%) completed the questionnaire. Appendix D provides a sample the study timeline.

**Vignette and Questionnaire Development.** A series of vignettes were developed that asked individuals to consider the effectiveness results from a recent education evaluation. The vignettes were structured to test the two study arms and two paradigms discussed earlier. The first study arm presented results from a high cost and easy to implement program. Two vignettes for this condition were developed to present frequentist and Bayesian paradigms, labeled conditions one and two. The second study arm presented results from a low cost and difficult to implement program. Two vignettes for this condition were developed to present frequentist and Bayesian paradigms, labeled conditions three and four. The vignettes included the appropriate p-value (Frequentist) or posterior probability (Bayesian) for this moderate level of evidence.<sup>4</sup> Appendix E presents the four different vignettes.

A survey questionnaire was developed to collect data to provide evidence for or against the aforementioned research questions. Respondents were asked to read two vignettes and complete two survey questionnaires. The items on each questionnaire were identical and respondents completed each questionnaire at the conclusion of each vignette. The survey questions measured their willingness to endorse the program based on the information from the previous scenario (i.e., effects under either one statistical paradigm, the cost information, and the implementation

---

<sup>4</sup> Prior research has shown individuals are more likely to endorse a program when results are presented under a Bayesian framework rather than a frequentist presentation for both the strong and moderate evidence scenarios with no differential endorsement for a weak evidence scenario (Chandler, Martinez, Finucane, Terziev, & Resch, 2019).

analysis information). Random assignment was used to determine study arms and order of the vignettes for each respondent. More specifically, respondents were randomly assigned to only one cost and implementation feasibility (i.e., low cost and difficult to implement versus high cost and easy to implement) study arm and one of two presentation orders (i.e., frequentist then Bayesian versus Bayesian then frequentist). The first group was low cost and difficult to implement with the frequentist presentation first and the Bayesian presentation second. The second group was low cost and difficult to implement with the Bayesian presentation first and the frequentist presentation second. The third group was high cost and easy to implement with the frequentist presentation first and the Bayesian presentation second. The final group was high cost and easy to implement with Bayesian presentation first and the frequentist presentation second. The resulting experimental design was a within subjects' design, with cost and implementation as a between-subjects factor and statistical scenario as a within-subjects factor (see also the Data Analytic Approach section of this Chapter).

The survey questionnaire took respondents approximately 5-10 minutes to complete. Endorsement of the program was measured through four distinct questions reflecting different types of endorsements. The first endorsement question asked the respondents to determine whether school officials are justified in endorsing the program for their school district. The second asked respondents to judge whether the local school board is justified in endorsing the program for their school. These questions are being asked since it does not require the respondent to make a decision whether they should encourage their Member to support the program, and the questions also provide an important distinction between key stakeholders in school

districts (i.e., superintendents—school officials and public school boards—typically, elected members of the community). The second pair of endorsement questions mirror the two types of endorsements legislative aides are often asked to recommend to their congressional superiors. The first asks whether or not a non-financial endorsement should be advanced for the program (e.g., a written letter from the Member in support of the program) and the second asks about a financial endorsement (e.g., where the Member casts a vote allocating federal dollars to the program).

Two additional questions were asked concerning whether respondents found the results of the study informative as well as the ease of interpretation. Respondents were also asked a series of demographic questions at the end of the survey. The survey was hosted on the Qualtrics online survey platform and informed consent was obtained prior to respondents beginning the survey. Respondents were randomly assigned to their study arm and vignette order using a random number generator within Qualtrics. Appendix F presents the survey questionnaire.

**Data Analytic Approach.** The vignettes in this study involve three varying conditions. These conditions are: (1) statistical framework (i.e., Bayesian versus frequentist), (2) cost (i.e., high cost versus low cost), and (3) implementation feasibility (easy to implement versus difficult to implement). Evidence strength (i.e., effect size and precision) is held constant at the moderate evidence level. Ideally, this model would be represented by three factors, each with two levels; in other words, a 2 x 2 x 2 ANOVA. However, I collapsed two dimensions of this model (i.e., cost and implementation feasibility) and did not test for conditions of high cost and difficult implementation or low cost and easy implementation. The reason for this decision is due to (a) sample size limitations and the need to maximize statistical power for a

small sample, (b) the need to limit response burden (i.e., only two scenarios could be presented to each respondent). A more thorough description of statistical power is discussed later in this methods section. Due to this methodological decision, the result is a 2 x 2 factorial repeated measures design.

The main dependent variable of interest (endorsement of the program) was measured on a continuous 7-point Likert scale under each of the four endorsement items. The end-point labels of the scale were “strongly disagree” and “strongly agree” with the middle value label being “neither agree nor disagree”. This scaling approach has been shown to reduce non-response bias (Courser & Lavrakas, 2012). The three aforementioned categorical independent variables were modeled (i.e., paradigm, scenario, and the interaction of paradigm and scenario). Since this study incorporates a continuous criterion variable with two binary predictor variables, a factorial repeated measures ANOVA, estimated under both a frequentist and a Bayesian framework, was the primary statistical model used to examine experimental effects.

**A Justification for Likert Scaling.** The use of a Likert enables more fluidity in response than a dichotomized yes versus no questionnaire. By employing a Likert scale, respondents are better able to frame their decision in terms of increasing degrees of either a favorable or unfavorable endorsement. Although decision making is ultimately binary, this research project was focused on understanding the scalar aspects of an agreeable versus disagreeable endorsement. Further, the flexibility of the Likert scale enables the responses to ultimately be recoded into binary yes versus no responses. Appendix G presents the analysis from a repeated measures logistic regression model based on binary responses (i.e., agree, strongly agree = 1; else 0).

**Statistical Model (Frequentist).** The ANOVA model and associated notation is presented below:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

Where:

$\mu$  = mean of all observations

$\mu + \alpha_j$  = mean of all observations under treatment  $j$  (statistical paradigm)

$\mu + \beta_k$  = mean of all observations under treatment  $k$  (high cost + easy implementation versus low cost + difficult implementation)

$\mu + \alpha_j + \beta_k + \alpha\beta_{jk}$  = mean of all observations under treatment  $j$  and  $k$

$\varepsilon_{ijk}$  = repeated-measures residual, with correlation  $\rho$  between observations within individuals

An a priori power analysis was completed for a 2 x 2 factorial repeated measures ANOVA design using a within subjects' design using the GPower 3.0 program (Faul, Erdfelder, Lang, & Buchner, 2007). Two-tailed  $p =$  values were employed with an alpha level of 0.05 and a power level of 0.80. An expected effect size  $f = .25$  was assumed, which is considered a medium effect size. Although the  $f$  effect size is an option for modeling effect sizes using an ANOVA model, the  $d$  effect size is a more common effect size metric. Given the prevalence of the  $d$  effect size over the  $f$  effect size, the online calculator Psychometrica was used to establish an equivalent Cohen's  $f$  to Cohen's  $d$ . The result was an  $f$  effect size of 0.25 equates to a  $d$  effect size of 0.5 (Lenhard & Lenhard, 2016). An effect size of  $d = 0.5$  is

appropriate because prior literature in this field has observed main interaction effect sizes above  $d = 0.5$  (Chandler, Martinez, Finucane, Terziev, & Resch, 2019).<sup>5</sup>

To model a main effect, the first factor (i.e., Bayesian versus frequentist) has two levels and hold one degree of freedom ( $2 - 1 = 1$ ). The second factor also has two levels (high cost and easy implementation versus low cost and difficult implementation) and retains one degree of freedom ( $2 - 1 = 1$ ). To estimate the degrees of freedom with the interaction effect  $(2 - 1) * (2 - 1) = 1$ df. Thus, the result is one degree of freedom for both main and interaction effects held across four distinct groups. Based on these conditions and an assumed correlation of .60 between repeated measures, a sample size of 28 respondents is necessary to achieve 80% power for detecting a main effect for the within-subjects factor, which in this study represents the effect of the frequentist versus Bayesian statistical paradigm.

**Statistical Model (Bayesian).** The Bayesian equivalence of the two-way ANOVA model is provided below (Kruschke, 2015)

$$P(\mu, \sigma | D) = \frac{P(D | \mu, \sigma) P(\mu, \sigma)}{\iint d\mu d\sigma P(D | \mu, \sigma) P(\mu, \sigma)}$$

One prior distribution was tested for the prior,  $P(\mu, \sigma)$  with an assumed normal distribution. Using the JASP Team (2020) statistical software program, a uniform prior distribution  $P(M)$  was assumed that assigned equal prior odds to all tested

---

<sup>5</sup> Prior research observed an  $F$ — value = 36.41 w/ approximately 231 subjects in each group (Chandler, Martinez, Finucane, Terziev, & Resch, 2019). In order to compute an expected  $d$  effect size from an  $F$ — value of 36.41, the formula  $d = \text{SQRT}((F * (n1 + n2) / (n1 * n2)))$  was used. This formula revealed an expected effect size of  $d = .56$  suggesting a value of  $d = .5$  is an obtainable effect size. The formula w/ numeric substitutions is presented:  $d = \text{SQRT}((36.41 * (231 + 231) / (231 * 231))) = .56$

models.  $BF_M$  was computed to measure the change odds from the prior to the posterior. Finally,  $BF_{10}$  was used to interpret the Bayes factor.

## **Chapter 4**

### **RESULTS**

The experimental results are presented in both frequentist and Bayesian frameworks. First, descriptive statistics are presented that describe both the Congressional staff sample and then the University of Delaware undergraduate sample. After the presentation of descriptive data, baseline equivalency tests are provided to verify the success of random assignment. Four separate chi-square tests were used for each of the Congressional staffer and undergraduate samples to assess baseline equivalency. The frequentist ANOVA model for impacts on target outcomes follows next and results are presented for both Congressional staffers and University of Delaware undergraduates. Finally, the Bayesian ANOVA equivalent is presented using a flat prior for both Congressional staffers and University of Delaware undergraduates. Bayes factor is used to provide interpretive guidance for the results. Bayes factor is useful for providing interpretive guidance because Bayes factor is a ratio of the likelihood of one hypotheses odds to the other hypothesis odds (“Bayes Factor: Simple Definition”). In sum, understanding the results in the context of the experimental hypotheses, the evidence enables the rejection of the null for the first hypothesis, but we fail to reject the null for hypotheses two and three.

Further analysis regarding a pooled sample of Congressional staffers and undergraduates is presented in the appendix along with histograms that visualize the distribution of data for the item level responses for both Congressional staffers and undergraduates. The visualizations are provided in appendix H and the pooled sample analysis in appendix I. Finally, appendix J provides information on the Benjamini-Hochberg procedure used for the false discovery rate adjustment.

## **Descriptive Statistics**

For the congressional staffer sample, descriptive statistics are presented for age, undergraduate major, number of completed statistics courses, and policy subject matter area focus. For the University of Delaware sample, descriptive statistics are presented for the year of current study, the undergraduate major in addition to Political Science, the number of completed statistics courses, and their policy subject matter area of interest. Thirty Congressional staffers completed the vignette experiment and thirty-six University of Delaware undergraduate students completed the experiment.

## **Congressional Staffers**

Table 1 presents the age of Congressional staffers. Nineteen Congressional staffers (63%) reported their age as 21-25 years, three Congressional staffers (10%) reported an age of 26-29, one Congressional staffer (3%) reported an age of 30-34, and five Congressional staffers (17%) reported an age of 40+. No staffers reported an age of 35-39 and two Congressional staffers (7%) did not report their age. These results are consistent with prior work that suggests most Congressional staffers are young professionals who are recent college graduates.

Table 1 Reported Age of Congressional Staffers

Age	N
21-25	19
26-29	3
30-34	1
40+	5
Total	30

Table 2 presents the undergraduate majors of the Congressional staffers. Nine different disciplines were reported for undergraduate majors with the most common reported major being Political Science. Nine Congressional staffers (30%) reported a major of Political Science. This is consistent with prior research suggesting Political Science is the most frequently selected major of Congressional staffers. Twenty Congressional staffers (67%) reported a major in the social sciences.

Table 2 Reported Undergraduate Majors of Congressional Staffers

Age	N
Biology	1
Business Administration	2
Criminal Justice	3
Economics	2
Education	1
History	5
Political Science	9
Public Health	1
Sociology	1
Did not report	5
Total	30

Table 3 presents the number of statistics courses completed by the Congressional Staffers. Nine Congressional staffers (30%) reported completing only

one course in statistics. Eleven Congressional staffers (37%) reported completing two courses in statistics. Two Congressional staffers (7%) reported completing three courses in statistics. Two Congressional staffers (7%) reported completing five or more courses in statistics. No Congressional staffers reported completing four courses in statistics. Four Congressional staffers (13%) reported completing no courses in statistics. Two Congressional staffers (7%) did not report the number of statistics courses completed. Interestingly, 19 Congressional staffers (63%) reported having completed only two courses in statistics suggesting the level of statistical knowledge obtained by staffers is minimal and one can reasonably surmise given this limited exposure to statistics that Congressional staffers are unfamiliar with Bayesian statistics.

Table 3      Reported Statistics Courses Completed by Congressional Staffers

Age	N
One course	9
Two courses	11
Three courses	2
Four courses	0
Five or more courses	2
No courses	4
Did not report	2
Total	30

Table 4 presents the policy subject area where Congressional staffers primarily work. Congressional staffers were asked to list their top three areas of policy work within their current policy portfolio. The item asked staffers to select up to three since many staffers work across policy subject matter areas. The area of defense policy was reported seven times (18%). The area of education policy was reported 12 times

(30%). The area of healthcare policy was reported 11 times (28%). The area of labor/employment policy was reported four times (10%). The area of housing/transportation was reported six times (15%). Ten Congressional staffers did not report at least one area of policy work. The data suggests the most frequently reported policy area for this sample of Congressional staffers is education.

Table 4      Reported Public Policy Work Area for Congressional Staffers

Age	N
Education	12
Healthcare	11
Labor/Employment	4
Housing/Transportation	6
Defense	7
Did not report	10
Total	50*

\*Respondents allowed to report up to three policy areas

### **University of Delaware Undergraduates**

Table 5 presents the current year of education for University of Delaware undergraduates. Ten undergraduates (28%) reported they were in their first year of study. Eight undergraduates (22%) reported they were in their second year of study. Eight undergraduates (22%) reported they were in their third year of study. Nine undergraduates (25%) reported they were in their fourth year of study. One undergraduate (3%) did not report their year of study. These results indicate an even distribution in terms of the current year of study for the undergraduate sample.

Table 5 Reported Year of Study for Undergraduates

Age	N
First year	10
Second year	8
Third year	8
Fourth year	9
Did not report	1
Total	36

Table 6 presents the undergraduate majors from the University of Delaware sample. All undergraduates were studying Political Science with another twelve different disciplines reported as a secondary area of study in addition to Political Science. Public Policy and Criminal Justice were the two most popular second disciplines reported with each having three undergraduates (16%) report the discipline. Ten undergraduates (53%) reported a second discipline in a social science.

Table 6 Reported Second Undergraduate Major for Undergraduates

Age	N
Business	1
Criminal Justice	5
Education	1
Environmental Studies	2
German	1
History	2
Marketing/Management	1
Psychology	2
Public Policy	3
Russian Studies	1
Sociology	1
Spanish	1
Total	21

Table 7 presents the number of statistics courses completed by the undergraduate sample. Twenty-two undergraduates (61%) reported completing only one course in statistics. One undergraduate (3%) reported completing two courses in statistics. Two undergraduates (6%) reported completing five or more courses in statistics. Ten undergraduates (28%) reported completing no courses in statistics. No undergraduates reported completing 3 or 4 courses in statistics. One undergraduate (3%) did not report the number of statistics courses completed. Notably, 32 undergraduates (89%) reported having completed none or only one course in statistics. The lack of statistical training on behalf of this sample is remarkable considering that 17 undergraduates (47%) are upperclassmen in this sample.

Table 7      Reported Statistics Courses Completed by Undergraduates

Age	N
One course	22
Two courses	1
Three courses	0
Four courses	0
Five or more courses	2
No courses	10
Did not report	1
Total	36

Table 8 presents the policy subject areas of interest for undergraduates. Undergraduates were asked to list their top three policy areas of interest. The area of defense policy was reported 18 times (18%). The area of education policy was reported 27 times (28%). The area of healthcare policy was reported 20 times (20%). The area of labor/employment policy was reported 19 times (19%). The area of

housing/transportation was reported 14 times (14%). One undergraduate did not report at least one area of policy subject matter interest. The data suggests the most frequently reported policy area of interest in this sample of undergraduates was education.

Table 8      Reported Public Policy Work Area of Interest for Undergraduates

Age	N
Education	27
Healthcare	20
Labor/Employment	19
Housing/Transportation	14
Defense	18
Did not report	1
Total	99*

\*Respondents allowed to report up to three policy areas

### **Baseline Tests of Equivalency**

Chi-square tests were used to understand the success of random assignment at establishing baseline equivalency. Four chi-square tests were conducted for each of the covariates leading to a total of eight tests for both the Congressional sample and University of Delaware undergraduate sample.

### **Congressional Staffers**

Four chi-square tests were conducted on the Congressional sample as baseline tests of equivalency to understand the effects of random assignment against the covariates of age, undergraduate major, number of completed statistics courses, and policy subject matter area focus. For the Congressional sample, six (20 percent)

respondents were randomized into condition one, eight (27 percent) respondents were randomized into condition two, nine (30 percent) respondents were randomized into condition three, and seven (23 percent) respondents were randomized into condition four. Tables 9 and 10 present the resultant p-values for each of the equivalency tests. No significant associations were found between assignment to condition and any covariates, suggesting random assignment was successful at establishing baseline equivalence.

Table 9 Baseline Equivalency Tests for Age and Undergraduate Major for Congressional Staffers

Variable	Age		Undergraduate Major	
	Chi-square	Sig (two-sided)	Chi-square	Sig (two-sided)
	12.23	0.43	56.29	0.50

Table 10 Baseline Equivalency Tests for Statistics Courses and Policy Area for Congressional Staffers

Variable	Statistics Courses		Policy Area	
	Chi-square	Sig (two-sided)	Chi-square	Sig (two-sided)
	15.96	0.39	46.08	0.12

### University of Delaware Undergraduates

Four chi-square tests were conducted on the University of Delaware undergraduate sample as baseline tests of equivalency to understand the effects of random assignment against the covariates of the year of current study, the undergraduate major in addition to Political Science, the number of completed

statistics courses, and their policy subject matter area of interest. For the undergraduate sample, eight (22 percent) respondents were randomized into condition one, 11 (31 percent) respondents were randomized into condition two, eight (22 percent) respondents were randomized into condition three, and nine (25 percent) respondents were randomized into condition four. Tables 11 and 12 present the resultant p-values for each of the equivalency tests. No significant associations were found between assignment to condition and any covariates, suggesting random assignment was successful at establishing baseline equivalence.

Table 11 Baseline Equivalency Tests for Age and Second Major for Undergraduates

Variable	Age		Second Major	
	Chi-square	Sig (two-sided)	Chi-square	Sig (two-sided)
	11.46	0.49	63.54	0.26

Table 12 Baseline Equivalency Tests for Statistics Courses and Policy Area for Undergraduates

Variable	Statistics Courses		Policy Area	
	Chi-square	Sig (two-sided)	Chi-square	Sig (two-sided)
	9.65	0.65	40.22	0.55

### Frequentist ANOVA Models

Six separate two-factor repeated-measures ANOVA frequentist models were conducted for both samples of Congressional staffers and University of Delaware undergraduates (i.e., one model for each of the six key survey response items). IBM Corp (2018) SPSS version 26.0, a computer software for conducting statistical

analyses was used to conduct the frequentist analysis. The dependent variable was the item level response and the factors tested included scenario (high cost/difficult to implement and low cost/easy to implement), paradigm (Bayesian versus frequentist), and an interaction for scenario and paradigm. Mean values, standard deviations, ANOVA f-values, and effect sizes are reported for the observed significant factor of paradigm. Significant results were not observed for the factors of scenario or the interaction factor for paradigm and scenario.

### **Congressional Staffers**

Table 13 presents the means, standard deviations, ANOVA f-values, and effect sizes for each of the items for the paradigm factor. All models revealed a statistically significant effect for the paradigm factor showing respondents are more likely to endorse a favorable view for the item if the vignette that preceded the item response was presented using the Bayesian paradigm. The main effect for scenario and the interaction of scenario and paradigm factors were not statistically significant in any of the models.

For the first item (i.e., “The results from this study are informative.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.003$ ) and effect size ( $d_{rm} = +0.72$ ). Congressional staffers were 3.4 times more likely to somewhat agree, agree, or strongly agree that “The results from this study are informative” under the Bayesian paradigm (see Appendix G for details). For the second item, (i.e., “The results from this study are easy to understand.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.015$ ) and effect size ( $d_{rm} = +0.71$ ). Congressional staffers were 2.2 times more

likely to somewhat agree, agree, or strongly agree that “The results from this study are easy to understand” under the Bayesian paradigm.

For the third item, (i.e., “Based on the results of this study, school officials are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.001$ ) and effect size ( $d_{rm} = +0.97$ ). Congressional staffers were 3.3 times more likely to somewhat agree, agree, or strongly agree that “school officials are justified in endorsing this program” under the Bayesian paradigm. For the fourth item, (i.e., “Based on the results of this study, members of local school boards are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.001$ ) and effect size ( $d_{rm} = +0.98$ ). Congressional staffers were 4.6 times more likely to somewhat agree, agree, or strongly agree that “members of local school boards are justified in endorsing this program” under the Bayesian paradigm.

For the fifth item, (i.e., “Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.003$ ) and effect size ( $d_{rm} = +1.16$ ). Congressional staffers were 2.2 times more likely to somewhat agree, agree, or strongly agree that “Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement” under the Bayesian paradigm. For the final item, (i.e., “Based on the results of this study, Congressional Members are justified in casting a vote, which

proves financial support for the program.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.002$ ) and effect size ( $d_{rm} = +1.27$ ). Congressional staffers were 2.3 times more likely to somewhat agree, agree, or strongly agree that “Congressional Members are justified in casting a vote, which proves financial support for the program” under the Bayesian paradigm.

Table 13 Overall Frequentist ANOVA Values for Each of the Items by Paradigm for the Congressional Staffers

Variable	Bayes		Frequentist		F	$d_{rm}$
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Item 1	5.41	0.90	4.67	1.56	10.43**	+0.72
Item 2	5.24	1.12	4.70	1.51	6.78*	+0.71
Item 3	4.90	1.29	3.83	1.34	12.50***	+0.97
Item 4	5.00	1.30	3.93	1.31	13.84***	+0.98
Item 5	4.76	1.35	4.03	1.56	10.38**	+1.16
Item 6	4.66	1.37	3.77	1.75	11.88**	+1.27

Note.  $d_{rm}$  is Cohen’s  $d$  for repeated measures. \*\*\* $p < 0.001$  \*\* $p < 0.01$  \* $p < 0.05$

### University of Delaware Undergraduates

Table 14 presents the mean, standard deviations, ANOVA f-values, and effect sizes for each of the items for the paradigm factor. Three models revealed a statistically significant effect for the paradigm factor, showing respondents are more likely to endorse a favorable view of the item if the vignette that precedes the item response was presented in the Bayesian paradigm. The main effect of scenario and the interaction of scenario and paradigm factors did not show any significant effects in any model.

For the first item (i.e., “The results from this study are informative.”), no statistically significant effect of paradigm was observed. For the second item (i.e., “The results from this study are easy to understand.”), no statistically significant effect of paradigm was observed.

For the third item (i.e., “Based on the results of this study, school officials are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.002$ ) and effect size ( $d_{rm} = +0.82$ ). Political Science undergraduates were 5.6 times more likely to somewhat agree, agree, or strongly agree that “school officials are justified in endorsing this program” under the Bayesian paradigm (see Appendix G for details). For the fourth item (i.e., “Based on the results of this study, members of local school boards are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.013$ ) and effect size ( $d_{rm} = +0.64$ ). Political Science undergraduates were 2.7 times more likely to somewhat agree, agree, or strongly agree that “school boards are justified in endorsing this program” under the Bayesian paradigm.

For the fifth item (i.e., “Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement.”), no statistically significant effect of paradigm was observed. For the final item (i.e., “Based on the results of this study, Congressional Members are justified in casting a vote, which proves financial support for the program.”), a statistically significant effect of paradigm was found, showing a more agreeable rating when the presented paradigm was Bayesian ( $p = 0.010$ ) and effect

size ( $d_{rm} = +0.75$ ). Political Science undergraduates were 3.3 times more likely to somewhat agree, agree, or strongly agree that “Congressional Members are justified in casting a vote, which proves financial support for the program” under the Bayesian paradigm.

Table 14 Overall Frequentist ANOVA Values for Each of the Items by Paradigm for the Undergraduates

Variable	Bayes		Frequentist		F	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Item 1	5.64	1.15	5.42	1.22	0.84	-----
Item 2	5.31	1.21	5.33	1.24	0.005	-----
Item 3	5.31	1.06	4.42	1.46	10.87**	+0.82
Item 4	5.36	1.15	4.67	1.49	6.87*	+0.64
Item 5	4.75	1.38	4.39	1.47	1.23	-----
Item 6	5.19	1.19	4.31	1.65	7.38*	+0.75

Note.  $d_{rm}$  is Cohen’s *d* for repeated measures. \*\*\* $p < 0.001$  \*\* $p < 0.01$  \* $p < 0.05$

### Bayesian ANOVA Model

Six separate two-factor repeated-measures Bayesian ANOVA models were conducted for both samples of Congressional staffers and University of Delaware undergraduates. JASP Team (2020) JASP version 0.12.2, a computer software for conducting statistical analyses was used to conduct the Bayesian analysis. The dependent variable was the item level response and the factors tested included scenario (high cost/difficult to implement and low cost/easy to implement), paradigm (Bayesian versus frequentist), and an interaction for scenario and paradigm. Bayesian ANOVA model results are presented with resulting values for  $P(M)$  which is the uniform distribution of the model’s prior probability,  $P(M|data)$  which shows the

posterior model's probability (i.e., the probability of the model given the data),  $BF_M$  which shows how the model has changed from the prior to the posterior,  $BF_{10}$  which is the Bayes' factor for the hypothesized model versus a null model, and the error percentage (Wagenmakers et al., 2018). Bayes factors (i.e.,  $BF_{10}$ ) are useful metrics for interpreting outcomes from Bayesian data analysis because of their widely agreed upon interpretive consensus. Essentially, the  $BF_{10}$  shows the ratio of the probability of the observed data under the null versus the alternative hypotheses (e.g.,  $p[\text{data}|H_a] / p[\text{data}|H_0]$ ). In other words, it shows how much more likely (or less likely) the data are under the alternative hypothesis.

There are eleven categories that govern the interpretation of Bayes factors (Lee & Wagenmakers, 2013). The first are factors  $> 100$ . The interpretation here is extreme evidence exists in support of the alternative hypothesis. The second are factors 30-100. The interpretation here is that very strong evidence exists in support of the alternative hypothesis. The third are factors 10-30. The interpretation here is that strong evidence exists in support of the alternative hypothesis. The fourth are factors 3-10. The interpretation here is that moderate evidence exists in support of the alternative hypothesis. The fifth are factors 1-3. The interpretation here is that anecdotal evidence exists for the alternative hypothesis. The sixth are factors of 1. The interpretation here is that no evidence exists in favor of either the alternative or null hypotheses. The seventh are factors  $1/3-1$ . The interpretation here is that anecdotal evidence exists in favor of the null hypothesis. The eighth are factors  $1/10-1/3$ . The interpretation here is that moderate evidence exists in favor of the null hypothesis. The ninth are factors  $1/30-1/10$ . The interpretation here is that strong evidence exists in favor of the null hypothesis. The tenth are factors  $1/100-1/30$ . The

interpretation here is that very strong evidence exists in favor of the null hypothesis.

The final are factors  $<1/100$ . The interpretation here is that extreme evidence exists in favor of the null hypothesis.

### **Congressional Staffers**

Table 15 presents the resultant prior probability ( $P(M)$ ), posterior probability ( $P(M|data)$ ), changes odds from prior to posterior ( $BF_M$ ), and the Bayes factor ( $BF_{10}$ ) for the Congressional sample. All ANOVA models demonstrated a change odds from prior to posterior ( $BF_M$ ) by an order of magnitude greater than three-fold in favor of the paradigm model. In other words, given the data, the prior odds of the model having changed as a result of paradigm is on the order of magnitude of at least three-fold compared to the change odds from the other models. The Bayes factors ( $BF_{10}$ ) indicate moderate evidence in favor of the alternative hypothesis for items one and two, strong evidence in favor of the alternative hypothesis for item five and six, and very strong evidence in favor of the alternative hypothesis for items three and four.

Table 15 Overall Bayesian ANOVA Values for Each of the Models for the Congressional Staffers

Variable	P(M)	P(M data)	$BF_M$	$BF_{10}$	error%
Item 1	0.200	0.581	5.554	7.941	1.244
Item 2	0.200	0.441	3.154	3.235	1.113
Item 3	0.200	0.652	7.488	57.613	0.780
Item 4	0.200	0.656	7.639	98.231	1.061
Item 5	0.200	0.551	4.902	15.250	0.986
Item 6	0.200	0.618	6.483	22.978	2.812

## University of Delaware Undergraduates

Table 16 presents the resultant prior probability ( $P(M)$ ), posterior probability ( $P(M|data)$ ), changes odds from prior to posterior ( $BF_M$ ), and the Bayes factor ( $BF_{10}$ ) for the undergraduate sample. Three ANOVA models demonstrated change odds from prior to posterior ( $BF_M$ ) by an order of magnitude of three fold or greater. In other words, given the data, the prior odds of the model having changed as a result of paradigm is greater in favor of the paradigm factor for at least three of the models. The Bayes factors ( $BF_{10}$ ) indicate moderate evidence in favor of the null hypothesis for item two, anecdotal evidence in favor of the null hypothesis for items one and five, moderate evidence in favor of the alternative hypothesis for items four and six, and very strong evidence in favor of the alternative hypothesis for item three.

Table 16 Overall Bayesian ANOVA Values for Each of the Models for the Undergraduates

Variable	$P(M)$	$P(M data)$	$BF_M$	$BF_{10}$	error%
Item 1	0.200	0.185	0.905	0.370	2.199
Item 2	0.200	0.122	0.557	0.242	1.163
Item 3	0.200	0.645	7.281	30.266	0.928
Item 4	0.200	0.465	3.473	4.934	0.847
Item 5	0.200	0.207	1.044	0.484	2.584
Item 6	0.200	0.463	3.453	8.904	1.160

## **Chapter 5**

### **DISCUSSION**

The results from this study, which seeks to understand how individuals' reason and make judgements when information is presented in frequentist versus Bayesian statistics, have strong implications for policy research and the social sciences more broadly. Results confirm that responses from Congressional staff and public policy undergraduates are more agreeable on several outcomes when research results are presented under a Bayesian paradigm. Metaphorically, these results confirm that the frequentist paradigm appears to convey a pessimistic, 'glass half empty' interpretation, while the Bayesian paradigm appears to convey an optimistic, 'glass half full' interpretation of statistical results from program impact evaluations.

More specifically, statistically significant results were observed for the Congressional sample revealing, when the preceding paradigm was Bayesian, a more agreeable response for items that asked whether the results presentation was informative and easy to understand. Although no statistically significant differences existed for these items with respect to the University of Delaware students this may be because Congressional staffers are better prepared and more experienced with reasoning about statistical information compared to the undergraduates in this sample. As observed in the descriptive results, which asked respondents to state the number of courses they have completed in statistics, nearly 90 percent of the undergraduate sample reported completing none to only one course in statistics, suggesting these students may not have enough training or experience in statistics to form an opinion regarding how they prefer to see results. On the other hand, Congressional staffers

who frequently review statistical information likely have thought more deeply about different presentations of statistical results on prior occasions.

Notably, items three and four, which ask if local school officials and then local school board members, are justified in endorsing the program, statistically significant results showing a more agreeable response when the results were presented in Bayesian terms were obtained for both the Congressional sample and the University of Delaware undergraduate sample. Most notably, item five that asks if Congressional Members are justified in lending their non-financial support for the program, and then item six, which asks if Congressional Members are justified in casting a vote that provides financial support for the program, both revealed statistically significant effects for the Congressional sample, suggesting a more agreeable response when the preceding paradigm was Bayesian. Similarly, item six revealed a statistically significant effect of paradigm for the University of Delaware undergraduate sample, but item five did not. Despite the lack of significance for items one, two, and five in the University of Delaware undergraduate sample, these results hold profound meaning for the field.

Congressional staff recognize the complexities that underlie a Members' vote. These complexities only become compounded when one considers votes that have financial implications. The observed outcome shows that Congressional staff are more likely to feel their Member is justified in voting for financial support of the program if the results of that program evaluation are presented in the Bayesian framework is tremendous. These results make strong contributions to the field and have profound implications, which should be considered by multiple stakeholders, including policy researchers, policymakers, and research methods instructors.

## **Contributions to Research**

Minimal prior literature exists that discusses differences in reasoning about frequentist and Bayesian statistics. Prior work had shown a more favorable result for information presented in Bayesian statistics when Amazon Turk recipients are asked to pretend that they are school district superintendents determining whether to endorse a new educational technology (Chandler, Martinez, Finucane, Terziev, & Resch, 2019). Interestingly, the results found in this study confirm this earlier observed result.

It is of strong interest that Congressional staffers voted to support a program when presented in Bayesian terms compared to those presented in frequentist terms for the endorsement of a program by school officials and school board members. Although this finding of local school endorsement is meaningful and consistent with prior literature, the more profound finding is that Congressional staffers feel their Member's vote for a program is more justified if the results of an evaluation for that program are presented in Bayesian terms. This finding is especially remarkable when one considers that decisions regarding whether a Member casts a vote for a program or policy are complex and often extend well beyond factors considered in these vignettes such as evidence strength, cost, and the feasibility of program implementation. One might even go as far and say that these decisions are indeed political. Despite the complexities that often arise regarding Members' actual decisions, when asked if Members would be justified in supporting the program both through non-financial as well as financial support, Congressional staffers overwhelmingly responded favorably when the presented paradigm was Bayesian. This has serious implications for policy decision making suggesting that a decision can be influenced by the statistical paradigm in which the resultant program evaluation is presented.

Prior research suggested that Bayesian statistics are preferred and easier to understand than frequentist statistics. Further, prior work examining adult decision making in the context of the Bayesian paradigm had also suggested that decision action in practice more closely represents those assumptions of the Bayesian framework (Pankin, 2013). Interestingly, the results observed here further confirm these findings. Next, I discuss the value of conducting research with Congressional staffers in the context of this project, especially considering that these Congressional staffers are a difficult population with which to conduct research.

Currently employed Congressional staffers were surveyed as part of this work. This one of a kind research study asked currently employed Congressional staffers to read statistical vignettes and then decide whether they would support the endorsement of the program or policy. Surveying Congressional staffers is notoriously difficult for several reasons. First, access to Congressional staffers is problematic. Capturing published lists of Congressional staffers' emails is difficult, and in-person recruitment of Congressional staffers is not feasible without access to Congressional offices in Washington D.C. or hundreds of local offices across the country. Further, many offices establish policies that prohibit staffers from replying to surveys and often do not make a distinction between research-based surveys and opinion polling surveys. Despite these challenges, this dissertation project was successful in obtaining informed consent and achieving a sufficient sample of Congressional staff to read the vignettes and complete the survey. Given the particular nature and relevance of the sample, there are several implications for the policy research community, policy makers, and University instructors who are preparing the next generation of statisticians and social scientists.

## **Implications for the Policy Research Community**

Policy researchers are the front-line social scientists tasked with the challenging work of ensuring complicated statistical information is presented in an easy to understand and potentially impactful manner. This work is high stakes, often involving the evaluation of programs or policies where many millions of dollars in federal funding are in play and used to improve outcomes for the most vulnerable members of society. For some researchers, grievances with how the frequentist paradigm presents results have surfaced. Frequentist statistics are thought to be a poor framework for presenting results from high profile studies because p-values are not well understood and often lead individuals to favor outcomes with significant p-values as is commonly observed with the p-value publication bias (Pitak-Arnop, Dhanuthia, Hemprich, & Pausch, 2010). Prior work by the American Statistical Association has attempted to disabuse these misconceptions by sharing guidance with researchers on how to contextualize p-values in terms of practical significance and highlighting the role other pieces of valuable statistical information such as effect sizes (American Statistical Association [ASA], “Statement on Statistical Significance and P-Values”). Notwithstanding this guidance, p-values are still considered by many to be unhelpful, leading many applied policy researchers to look towards other approaches, such as the Bayesian paradigm.

The results of this dissertation suggest the Bayesian paradigm may indeed be more useful and impactful for sharing results with Congressional staffers. Accepting that Bayesian statistics present a more supportive framework for policy action has strong implications for policy researchers, including those dedicated to a wide range of activities that use research evidence to inform curriculum in schools and other areas. As a result, I suggest three immediate action steps policy researchers and

methodologists should take to ensure they are prepared to use Bayesian statistics appropriately in the context of policy research. Policy researchers must reach consensus on aspects of applied Bayesian statistics if they are to use these frameworks successfully with respect to policy action. These include consensus on when conditional versus unconditional priors are used, consensus on the presentation of Bayesian results, and guidelines for interpreting Bayesian results.

First, guidelines must be established on when to use conditional versus unconditional priors. Although some broad suggestions exist within academic circles, codified guidelines by governing statistical bodies should establish appropriate situations where conditional priors can be used and other situations where unconditional priors should be used. This is especially important since the use of conditional priors are typically set by subject matter experts or derived from prior literature. Second, guidelines ought to be established on how to present Bayesian results. In terms of presentation, what statistical information will supplement the posterior probability? Will effect sizes, credible intervals, and regions of practical interest all be provided? Assessing tradeoffs between the presentation of these different pieces of statistical information and what types of visual aides are most helpful is a necessity. Finally, guidelines for how to interpret Bayesian results in the context of policy findings should be established. Although some guidelines have already been suggested (for example, the Bayes factor used in this dissertation), gaining consensus on interpretation ought to be produced and shared with the community. If consensus guidelines are not shared, then one risks inconsistency in interpretation and weak evidence may be endorsed. Consensus across these areas and perhaps other areas is important as policy researchers move towards an era where

Bayesian statistics proliferate over the frequentist. It is consensus only that will ensure high quality scientific standards and trust by the public for the scientific process.

### **Implications for Congressional Staffers**

The presentation of results in the Bayesian framework is likely not to be a sweeping change that happens in a few weeks, months, or maybe even not for years. The change is likely to be gradual and so for the immediate future, I argue policy makers are likely to continue to be asked to pass judgment on policies informed by research presented in frequentist results. As such, policy makers must exercise caution when reviewing results presented in frequentist terms. For as this dissertation has demonstrated, policy makers are subject to being less likely to endorse policies or programs when research in support of those policies or programs is presented in frequentist terms.

One strategy Congressional staffers can use is to be proactive about their outreach. Many private research organizations who conduct policy work often have Government Affairs Departments that can connect the Congressional staffer with the researcher who conducted the work. In fact, some of these organizations even engage in proactive outreach. Congressional staffers are encouraged to speak with authors of policy briefs to help interpret findings. Interpreting frequentist findings is not an easy task and as this dissertation has shown, Congressional staffers are prone to not support a policy or program despite optimal conditions that would favor such support. Understanding the interpretive limits of the p-value and how the p-value can be interpreted within the context of effect sizes and other pieces of statistical information is essential for Congressional staffers seeking to make the best decision action.

Further knowing that Bayesian statistics are likely to take more prominence in the coming years, Congressional staffers should prepare for the eventual emergence of Bayesian statistics as the common language for the presentation of policy results.

Bayesian statistics is likely to play a more prominent role in policy research in the coming decades, and Congressional staffers can take steps to prepare. Congressional staffers can engage in self-education, whether that be through informal study of Bayesian statistics or more formal coursework. This education will help Congressional staffers to have better grasp of Bayesian statistical concepts and application of those concepts. Further, Congressional staffers could initiate a call for action, most likely through federal research funding agencies, regarding guidelines for how Bayesian statistical information is presented to their community. As I argued for earlier, the policy research community should already be making strides to establish guidelines for the best practices related to presenting Bayesian statistical findings to Congressional staffers. By working in concordance with policy researchers and federal funding agencies, Congressional staffers and policy researchers can set the direction for how Bayesian statistics will take their place in the policy research community.

### **Implications for University Instructors Teaching Statistics**

Important implications exist for instructors of University statistics courses. As I discussed earlier in this dissertation, there still exists only limited opportunities to study Bayesian statistics. Although classes exist in mathematics and statistics departments throughout many colleges and universities, the proliferation of Bayesian statistics in the social sciences is new and still emerging. In fact, as I discussed earlier, the University of Delaware did not offer a course in applied Bayesian analysis until the spring 2018. This profound point should raise awareness across universities that

more work is needed to prepare the next generation of social scientists to conduct analyses using Bayesian statistics. These implications extend beyond just one course, and much more careful analysis of education curriculum is needed if universities are to meet the future demand for researchers trained in Bayesian analysis.

The foundation of education in statistics is still far too dominated by the frequentist paradigm. One class in applied Bayesian analysis for advanced social science graduate students is insufficient to prepare the next generation of researchers. For almost every statistical test in frequentist terms there is a comparative Bayesian equivalent. Parametric and non-parametric tests exist in the Bayesian world similar to the frequentist world. ANOVA and linear regression also exist in the Bayesian world. More advanced statistical concepts such as hierarchical linear modeling, structural equation modeling, and factor analytic models also have Bayesian siblings. Hence, I argue that Universities have a tremendous amount of work ahead of them—faculty need to fundamentally examine how they teach students statistics from the basics although through to the most advanced courses. Before this chapter ends with a discussion acknowledging limitations of this dissertation, it is important that we revisit the fundamental differences between frequentist and Bayesian statistics to serve as a reminder that this dissertation has implications beyond policy research, cognitive science, and decision action.

### **Beyond Cognitive Science: Fundamental Ontological Differences**

The empirical nature of this dissertation combined with its orientation to the field of decision action should not deemphasize the fundamental implications this work has for the field of applied statistics. Although there is value in understanding the results of this dissertation for the field of decision action, ultimately one must not

forget to underscore the implications this work has more broadly for the field of statistical inference and the larger social sciences. There are times in the history of science where the reconciliation of two worldviews within a discipline is simply not possible. The field of statistics is not immune from these types of irreconcilable events, and it is my contention, that we are faced with this reality with respect to the frequentist and Bayesian paradigms. Frequentist and Bayesian statistics can coexist as two statistical paradigms but ultimately, cannot be reconciled. These paradigms represent two fundamentally different ontological stances with deep implications that strike to the core of scientific inference. Earlier in this dissertation a thorough review of the differences between the frequentist and Bayesian paradigms was provided in order to highlight the ontological implications of both positions. I revisit these distinctions to reinforce the notion that these statistical paradigms represent fundamentally different frameworks and it is these differences that have strong implications for the field of statistics and the broader social sciences.

The frequentist statistician is sometimes metaphorically equated with a pessimist. The reason that the frequentist receives the pessimist label is because the frequentist acknowledges inherent limits to scientific conclusions believing that our conclusions can never truly capture the world and all its mystery. Instead, our conclusions represent the closest approximation to truth. As such, the statistical formulations of the frequentist generate a parameter that can never truly be known but is suggested to be fixed. This fixed parameter has a fluid confidence interval that is subject to change with each statistical test. As a logical consequence of this framework, the frequentist is interested in replication of findings because it is replication that builds assurances that the convergence of multiple similar findings

leads to closer and closer approximations. This worldview is fundamentally different than that of the Bayesian.

The Bayesian statistician is sometimes metaphorically equated with an optimist. The Bayesian, however, is not a blind optimist who refuses to acknowledge the limits of science but to the contrary, quite honestly acknowledges the inherent limits to scientific conclusions. Unlike the frequentist though, the Bayesian does not believe simple replication is the solution. Instead the Bayesian establishes a framework of conditional probabilities where the parameter is fluid with fixed credible intervals containing the parameter. As a consequence of this framework, Bayesians are not interested in simply replicating their results, and instead they seek to update their results as more information becomes available. This notion of updating results as opposed to replicating results is a fundamental difference of the Bayesian paradigm. This fundamental difference has strong implications for the field of statistics and the broader social sciences.

Replication is fundamental to science. Replication is one of the hallmarks that separates science from other bodies of knowledge that is why the infamous replication crisis (Maxwell, Lau, & Howard, 2015) posed such a terrible threat to science as a body of knowledge. Recognizing this fundamental nature, it is not that the Bayesian is uninterested in replication but rather the Bayesian allows for replicated results to be incorporated into more robust future estimates. These fundamental views of fixed versus fluctuating intervals and a fixed versus fluctuating parameter must not be lost in a dialogue of frequentist versus Bayesian statistics. These concepts are critical to the dialogue, and they show us that whichever paradigm ultimately dominates the next

century of research, every researcher who selects either paradigm is responsible for acknowledging these underlying differences.

### **Limitations**

There are a few limitations with respect to this study. First, the separate effects of cost and implementation cannot be disentangled. A fundamental decision of this dissertation was not to test the obvious scenarios. In other words, it was hypothesized that a study condition that described a policy that was ultimately low cost and easy to implement should not be tested empirically since undoubtedly such a program would receive support. Similarly, it was hypothesized that a study condition that described a policy that was both high cost and difficult to implement should not be tested empirically since it is likely such a program would not receive support. Although this decision did not lack reasonable justification, nonetheless the inability to distinguish between policy conditions that are both low cost and easy to implement as well as high cost and difficult to implement presents as a limitation to this dissertation. With respect to the current experimental set-up the tested scenarios remained as a between subjects' factor. In other words, participants saw either high cost and easy to implement or low cost and difficult to implement. Since the factor scenario was a between subjects' factor the statistical power for this factor is inadequate to observe differences and thus serves as a second limitation to this study.

Statistical inference involves many components of statistical information beyond p-values, posterior probabilities, and effect sizes. Although some may argue that these pieces of statistical information are essential to decision making, additional pieces of information such as the frequentist confidence interval and Bayesian credible interval are also relevant for decision makers. Others may argue that disclosing the

region of practical interest or conducting tests of equivalence are equally necessary components for decision making. It is not my contention that these pieces of information are not relevant and as such, a third limitation of this dissertation is that these varied features of statistical results were not presented in the vignettes. It may be the case that these elements influence decision making when offered. Finally, despite the high level of causal inference given that this study utilized a true experimental design, it was nonetheless a first of its kind study with this specific population of Congressional aides. As replication is the hallmark of scientific inquiry, further studies should attempt to update and extend the findings discovered through this investigation in favor of the Bayesian paradigm, as additional data and information can be used to add detail and improve precision of our understanding of the implications of Bayesian methods.

## **Chapter 6**

### **CONCLUSION**

This dissertation was an attempt to break new ground in the field investigating how individuals make judgements about statistical information. Although the implications of this dissertation are evident for the fields of cognitive science and decision action, the results from this dissertation are also highly relevant for applied statisticians determining best practices for the presentation of results. Drawing upon a diverse array of literature, this dissertation argued that individuals generally find probability statements more intuitive than p-values and that a conditional framework for analyzing data is more useful than the unconditional framework posed by the frequentist.

Prior empirical work suggested that individuals find Bayesian results easier to understand and are more likely to positively endorse a technology when evidence concerning that technology is presented in a Bayesian context. This dissertation sought to extend that early work by examining core elements regarding decision action that go beyond evidence strength. Recognizing that elements such as cost and implementation feasibility are central to any practical decision, this dissertation sought to understand the role Bayesian versus frequentist statistics has in influencing judgement about an education program when the evidence strength for that program is held constant and practical elements like program cost and implementation feasibility are involved. A within subjects' vignette experiment was designed to test the hypothesis that despite varying across conditions of cost and implementation feasibility, the Bayesian conditions would produce more positive endorsement compared to their frequentist counterparts.

The ambitious population recruited for this sample included 30 Congressional staffers and 36 University of Delaware undergraduate Political Science students. In a first of its kind study, currently employed Congressional staffers completed the vignette experiment and reported their willingness to endorse an education program when results are presented under a Bayesian versus a frequentist paradigm. Two factor mixed ANOVA models were produced to analyze the data and suggested that the Bayesian framework is found to be easier to understand and more informative for the presentation of results from a policy evaluation. Further, respondents were more likely to endorse a positive affirmation for the justification of support of the policy by local school board members as well as local school officials, and even a non-financial as well as a financial vote of endorsement by their Congressional Member when results of the study are presented in Bayesian terms. These results have profound effects for the field and the implications argued for here suggest several steps various stakeholder groups might take to ensure policy soundness of decisions given the observed findings.

The resultant analysis was also conducted in the Bayesian paradigm. Bayes factors were produced for the Bayesian ANOVA. All ANOVA models produced showed support for the alternative hypothesis although the support varied substantially within the established thresholds for interpreting Bayes factors. Further, the interpretation of the Bayesian ANOVA models does not necessarily align easily with the frequentist ANOVA models further highlighting the differences in interpretation between these paradigms. The results also highlight the importance of the prior in Bayesian analysis and with further research in this field, more established priors can be produced to inform more robust posterior estimates.

The title of this dissertation addressed the issue of the frequentist versus Bayesian paradigm with the metaphorical notion of a pessimist versus optimist. It is this author's belief that applied statistics are at a fundamental transformative point and the predicted outcome is truly a matter of conjecture whether the conclusion will be optimistic or pessimistic. In my view, Bayesian statistics is a more intuitive framework and can be better applied for practical decision making than the frequentist framework. As mentioned earlier, the popularization of p-values in contemporary statistics was never meant for p-values to be used in the manner that they have been used for so many years. Unfortunately, too many decision actors have made the incorrect assertion that if the p-value is significant then the program survives, and if the p-value is not significant then the program is eliminated. This type of binary decision making is not helpful for practical decisions and instead, a conditional framework that allows information to be updated as new information is obtained provides a more useful framework. Ultimately, guidelines need to be established for how to utilize Bayesian frameworks, and both producers of Bayesian analytics (e.g., applied statisticians) and consumers of Bayesian analytics (e.g., Congressional staffers) need more education and training in the use of these techniques.

In conclusion, I end my dissertation with an optimistic view of the future. I believe if training and education is appropriately offered, then the future of Bayesian analytics is bright, and as researchers, we will welcome a brave new world of applied statistical inquiry where our estimates have the freedom to be updated as more information becomes available, ultimately leading to a more secure version of truth. However, if we fail to establish guidelines and our universities fail to offer adequate training on these methods, then I worry we risk wasted years where Bayesian analytics

are used by groups who do not understand how to use them properly nor how to interpret them. Ultimately, pessimism versus optimism is a state of mind and a matter of true subjective impression. Hence, I conclude with the glass half full mentality confident that my colleagues will usher in a bright future for Bayesian producers and consumers of analytics. It is my professional and humbling pleasure to contribute but a small piece to this dialogue.

## REFERENCES

- American Psychological Association. (2018). Hypothesis. Retrieved from <https://dictionary.apa.org/hypothesis>
- American Psychological Association. (2018). Experiment. Retrieved from <https://dictionary.apa.org/experiment>
- American Statistical Association. (2016). Statement on statistical significance and p-values. Retrieved from <https://www.amstat.org/#>
- Andersen, H., & Hepburn, B. (2015). Scientific method. *The Stanford Encyclopedia of Philosophy*. Retrieved from <https://plato.stanford.edu/entries/scientific-method/>
- Bayes Factor: Simple Definition. (2018). Retrieved from <https://www.statisticshowto.com/bayes-factor-definition/>
- Bayesian vs Frequentist Approach: Same Data, Opposite Results (n.d.). Retrieved from <https://365datascience.com/bayesian-vs-frequentist-approach/>
- Buchinsky, F. J., & Chadha, N. K. (2017). To p or not to p: Backing bayesian statistics. *Otolaryngology—Head and Neck Surgery*, 157, 915-918. doi: [10.1177/0194599817739260](https://doi.org/10.1177/0194599817739260)
- Chandler, J.J., Martinez, I., Finucane, M. M., Terziev, J. G., & Resch, A. M. (2019). Speaking on data's behalf: What researchers say and how audiences choose. *Evaluation Review*. <https://doi.org/10.1177/0193841X19834968>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.
- Courser, M. W., & Lavrakas, P. J. (2012). Item-nonresponse and the 10-point response scale in telephone surveys. *Survey Practice*, 5, 1-7. doi: [10.29115/SP-2012-0021](https://doi.org/10.29115/SP-2012-0021)
- Dantzig, G. B., & Wald, A. (1951). On the fundamental lemma of neyman and pearson. *The Annals of Mathematical Statistics*, 22, 87-93. Retrieved from <https://projecteuclid.org/euclid.aoms/1177729695>
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274-290. doi: [10.1177/1745691611406920](https://doi.org/10.1177/1745691611406920)

- Efron, B. (1986). Why isn't everyone a bayesian? *The American Statistician*, 40, 1-5. [doi: 10.1080/00031305.1986.10475342](https://doi.org/10.1080/00031305.1986.10475342)
- Epstein, L., Martin, A. D., & Schneider, M. M. (2006). On the effective communication of the results of empirical studies, part 1. *Vanderbilt Law Review*, 59, 1811-1871. Retrieved from <https://deepblue.lib.umich.edu/handle/2027.42/116221>
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2017). How to become a bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25, 219-234. [doi: 10.3758/s13423-017-1317-5](https://doi.org/10.3758/s13423-017-1317-5)
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149-1160. [doi:10.3758/BRM.41.4.1149](https://doi.org/10.3758/BRM.41.4.1149)
- Fisher, R. (1954). *Statistical methods for research workers*. Kalpaz Publications.
- Gelman, A. (2008). Objections to bayesian statistics. *Bayesian Analysis*, 3, 445-450. [doi:10.1214/08-BA318](https://doi.org/10.1214/08-BA318)
- Gelman, A. (2008). Teaching bayes to graduate students in political science, sociology, public health, education, economics. *The American Statistician*, 62, 202-205. [doi: 10.1198/000313008X330829](https://doi.org/10.1198/000313008X330829)
- Glen, S. (2015). "Benjamini-Hochberg Procedure". Retrieved from <https://www.statisticshowto.com/benjamini-hochberg-procedure/>
- Gliner, J. A., Morgan, G. A., Leech, N. L., & Harmon, R. J. (2001). Problems with null hypothesis significance testing. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40, 250-252. <https://doi.org/10.1097/00004583-200102000-00021>
- Gotzsche, P.C., & Loannidis, J. P. A. (2012). Content area experts as authors: Helpful or harmful for systematic reviews and meta-analyses?. *BMJ*, 345, e7031. [doi: https://doi.org/10.1136/bmj.e7031](https://doi.org/10.1136/bmj.e7031)
- Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138, 1085-1108. [doi:10.1037/a0028044](https://doi.org/10.1037/a0028044)
- Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *WIREs Cognitive Science*, 6, 75-86. [doi: 10.1002/wcs.1330](https://doi.org/10.1002/wcs.1330)

- Gurrin, L. C., Kurinczuk, J. J., & Burton, P. R. (2000). Bayesian statistics in medical research: An intuitive alternative to conventional data analysis. *Journal of Evaluation in Clinical Practice*, 6, 193-204. <https://doi.org/10.1046/j.1365-2753.2000.00216.x>
- Hamra, G., MacLehose, R., & Richardson, D. (2013). Markov chain monte carlo: An introduction for epidemiologists. *International Journal of Epidemiology*, 42, 627-634. doi: [10.1093/ije/dyt043](https://doi.org/10.1093/ije/dyt043)
- Head, B. W. (2008). Three lenses of evidence-based policy. *Australian Journal of Public Administration*, 67, 1-11. <https://doi.org/10.1111/j.1467-8500.2007.00564.x>
- Hoekstra, R., Morey, R.D., Rouder, J.N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157-1164. doi: [10.3758/s13423-013-0572-3](https://doi.org/10.3758/s13423-013-0572-3)
- Hollands, F. M., & Levin, H. M. (2017). The critical importance of costs for education decisions (REL 2017–274). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- IBM Corp. (2018). *IBM SPSS Statistics for Windows* (Version 26.0). Armonk, NY: IBM Corp. <https://www.ibm.com/analytics/spss-statistics-software>
- JASP Team. (2020). *JASP* (Version 0.12.2). <https://jasp-stats.org/>
- James, J. (2019). Bayes's theorem. *The Stanford Encyclopedia of Philosophy*. Retrieved from <https://plato.stanford.edu/entries/bayes-theorem/>
- Jordan, M. I. (2009). *Are you a bayesian or a frequentist* [PDF document]. Retrieved from Lecture Notes Online Website: [http://videolectures.net/mlss09uk\\_jordan\\_bfway/](http://videolectures.net/mlss09uk_jordan_bfway/)
- Kaplan, D. (2018). Bayesian inference for social policy research (OPRE Report 2019 - 36). Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from [https://www.acf.hhs.gov/sites/default/files/opre/opre\\_kaplanbrief\\_021419\\_508.pdf](https://www.acf.hhs.gov/sites/default/files/opre/opre_kaplanbrief_021419_508.pdf)
- Khun, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

- Kruschke, J. K. (2015). *Doing bayesian data analysis (second edition)*. London, UK: Academic Press.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8, 355-362. doi: [10.1177/1948550617697177](https://doi.org/10.1177/1948550617697177)
- Lecoutre, B. (2006). Training students and researchers in bayesian methods. *Journal of Data Science*, 4, 207-232. Retrieved from <http://www.jds-online.com/v4-2>
- Lee, M.D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge, UK: Cambridge University Press.
- Lenhard, W., & Lenhard, A. (2016). Calculation of Effect Sizes. *Psychometrica*. doi: [10.13140/RG.2.1.3478.4245](https://doi.org/10.13140/RG.2.1.3478.4245)
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15, 22-25. <https://doi.org/10.1111/j.1467-9639.1993.tb00252.x>
- Malakoff, D. (1999). Bayes offers a ‘new’ way to make sense of numbers. *Science*, 286, 1460-1464. doi: [10.1126/science.286.5444.1460](https://doi.org/10.1126/science.286.5444.1460)
- Marken, S., & Auter, Z. (2017). What are the best incentives for web surveys? Retrieved from <https://news.gallup.com/opinion/methodology/224216/best-incentives-web-surveys.aspx>
- Maxwell, S.E., Lau, M.Y., & Howard, G.S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70, 487-498. doi: [10.1037/a0039400](https://doi.org/10.1037/a0039400)
- May, H. (2004) Making statistics more meaningful for policy research and program evaluation. *American Journal of Evaluation*, 25(4), 525-540. <https://doi.org/10.1177/109821400402500408>
- Maynard, R. A. (2018). The role of federal agencies in creating and administering evidence-based policies. *The Annals of the American Academy of Political and Social Science*, 678, 134-144. <https://doi.org/10.1177/0002716218768742>
- Montori, V. M., & Guyatt, G. H. (2008). Progress in evidence-based medicine. *JAMA Classics*, 300, 1814-1816. doi:[10.1001/jama.300.15.1814](https://doi.org/10.1001/jama.300.15.1814)

- Montgomery, J. M., & Nyhan, B. (2017). The effects of congressional staff networks in the US house of representatives. *The Journal of Politics*, 79, 745-761. <https://doi.org/10.1086/690301>
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, 82, 591-605. [doi:10.1111/j.1469-185X.2007.00027.x](https://doi.org/10.1111/j.1469-185X.2007.00027.x)
- Orloff, J., & Bloom, J. (2014). *Comparison of frequentist and bayesian inference*. [PDF document]. Retrieved from Lecture Notes Online Website: [https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18\\_05S14\\_Reading20.pdf](https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading20.pdf)
- Paap, R. (2001). What are the advantages of MCMC based inference in latent variable models? *Statistica Neerlandica*, 56, 2-22. <https://doi.org/10.1111/1467-9574.00060>
- Panagiotakos, D. B. (2008). The value of p-value in biomedical research. *The Open Cardiovascular Medicine Journal*, 2, 97-99. [doi:10.2174/1874192400802010097](https://doi.org/10.2174/1874192400802010097)
- Pankin, J. (2013). *Schema theory* [PDF document]. Retrieved from Lecture Notes Online Website: [http://web.mit.edu/pankin/www/Schema\\_Theory\\_and\\_Concept\\_Formation.pdf](http://web.mit.edu/pankin/www/Schema_Theory_and_Concept_Formation.pdf)
- Parker, M. (2004). *Foundations of statistics—frequentist and bayesian* [PDF document]. Retrieved from Lecture Notes Online Website: [http://www.austincc.edu/mparker/stat/nov04/talk\\_nov04.pdf](http://www.austincc.edu/mparker/stat/nov04/talk_nov04.pdf)
- Pigliucci, M. (2009). Hypotheses? Forget about it! *Philosophy Now*. Retrieved from [https://philosophynow.org/issues/74/Hypotheses\\_Forget\\_About\\_It](https://philosophynow.org/issues/74/Hypotheses_Forget_About_It)
- Pitak-Arnop, P., Dhanuthai, K., Hemprich, A., & Pausch, N. C. (2010). Misleading p-value: Do you recognize it? *European Journal of Dentistry*, 4, 356-358. [doi:10.1055/s-0039-1697852](https://doi.org/10.1055/s-0039-1697852)
- Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). Joining forces of bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, 371. <https://doi.org/10.1098/rsta.2011.0544>

- Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to markov chain monte—carlo sampling. *Psyehonomic Bulletin & Review*, 25, 143-154. <https://doi.org/10.3758/s13423-016-1015-8>
- R Core Team (2013). R: A language and envrionemnt for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To bayes or not to bayes, from whether to when: Applications of bayesian methodology to modeling. *Structural Equation Modeling*, 11, 424-451. doi: [10.1207/s15328007sem1103\\_7](https://doi.org/10.1207/s15328007sem1103_7)
- Sharot, T. (2011). The optimism bias. *Current Biology*, 21, 941-945. <https://doi.org/10.1016/j.cub.2011.10.030>
- Silva, I. R. (2017). On the correspondence between frequentist and bayesian rests. *Communications in Statistics—Theory and Methods*, 47, 3477-3487. doi: [10.1080/03610926.2017.1359296](https://doi.org/10.1080/03610926.2017.1359296)
- Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *ANNALS of the American Academy of Political Science*, 645, 112-141. <https://doi.org/10.1177/0002716212458082>
- Social Experiment. (n.d.). Retrieved from <https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/social-experiment#A>
- Stahl, S., & Johnson, P. E. (2007). *Understanding modern mathematics*. Sudbury, MA: Jones and Bartlett Publishers.
- Stone, D., & Denham, A. (2004). Think tank traditions: Policy research and the politics of ideas. *Public Administration*, 84, 1085-1114. [https://doi.org/10.1111/j.1467-9299.2006.00628\\_7.x](https://doi.org/10.1111/j.1467-9299.2006.00628_7.x)
- The Cohen’s d Formula (n.d.). Retrieved from <https://trendingsideways.com/the-cohens-d-formula>
- The Normal Distribution. (n.d.). Retrieved from <https://stattrek.com/probability-distributions/normal.aspx>
- Tolchin, M. (1991, November 12). Congress’s influential aides discover power but little glory on capitol hill. *The New York Times*. Retrieved from <https://www.nytimes.com/>

- Trafimow, D. (2018, June 15). The p-value ban, revisited: An interview with David Trafimow, PhD Basic and Applied Social Psychology (White, H.). Retrieved from <http://explore.tandfonline.com/page/beh/hbas-editor-interview-2018>
- Vose. (2017). Subjective priors. Retrieved from <https://www.vosesoftware.com/riskwiki/Subjectivepriors.php>
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kerstern, E.-J., van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with jasp. *Psychonomic Bulletin and Review*, 25, 58-76. doi: [10.3758/s13423-017-1323-7](https://doi.org/10.3758/s13423-017-1323-7)
- Wilcox, R. R., & Serang, S. (2017). Hypothesis testing, p values, confidence intervals, measures of effect size, and bayesian methods in light of modern robust techniques. *Educational and Psychological Measurement*, 77, 673-689. doi: [10.1177/0013164416667983](https://doi.org/10.1177/0013164416667983)
- Wilson, G. (2003). Tides of change: Is bayesian the new paradigm in statistics. *Journal of Statistical Planning and Inference*, 113, 371-374. [https://doi.org/10.1016/S0378-3758\(01\)00306-8](https://doi.org/10.1016/S0378-3758(01)00306-8)
- Working as a Legislative Assistant (n.d.). Retrieved from <https://www.zippia.com/legislative-assistant-jobs/>

## Appendix A

### SAMPLE EMAIL FOR RECRUITMENT OF LEGISLATIVE AIDES

Dear [Insert First Name],

I am writing to ask for your help in participation of a study investigating how individuals make judgements about statistical information. If you agree to participate, you will be asked to read one statistical vignette and answer a set of short survey questions. Reading the vignette and answering the short survey questionnaire should not take longer than 10 minutes.

I am conducting this research as part of my PhD studies at the University of Delaware. Your participation in this study will be a tremendous help towards my dissertation project. Further, you will be participating in important research which seeks to understand how individuals like yourself ultimately make judgments about statistical information. The goal of this research is to inform public policy researchers on the appropriate procedures for presenting statistical information on evaluations of public policies and programs to legislative aides like yourself. Only you can help inform this important research and without your participation, this research will not happen.

As a token of appreciation for participating in the survey, you will receive a \$10 Amazon gift card and be entered into a raffle to receive a \$35 Amazon gift card. I appreciate your consideration and you can contact me with any questions at [ahurwitz@udel.edu](mailto:ahurwitz@udel.edu).

Sincerely,

Andrew Hurwitz

## **Appendix B**

### **SAMPLE EMAIL TO INSTRUCTORS OF POLITICAL SCIENCE UNDERGRADUATES**

Dear Professor [Insert Last Name],

I am writing to ask for help in recruiting your students to participate in a study investigating how individuals make judgements about statistical information. The study asks students to read one statistical vignette and answer a set of short survey questions. Reading the vignette and answering the short survey questionnaire should not take longer than 10 minutes.

I am conducting this research as part of my PhD studies at the University of Delaware and request your permission to come before your class at a date and time you specify to make a recruitment speech directly to your students. The goal of my research is to inform applied social scientists on the differences of interpretation when individuals are presented with statistical information under a Bayesian versus Frequentist paradigm. I am seeking to recruit a sample of Political Science undergraduates since they are most likely to represent actual legislative aides.

Your support for recruiting students to participate in my dissertation project would be a tremendous help. If you could please respond to this message and indicate whether or not I have your support and the date and time you prefer me to arrive to your class I would greatly appreciate it. I am happy to share my recruitment speech at the beginning or end of your class and will take less than five minutes of your class time.

Most respectfully,

Andrew Hurwitz

## **Appendix C**

### **SAMPLE RECRUITMENT SPEECH TO POLITICAL SCIENCE UNDERGRADUATES**

Hello! My name is Andrew, and your instructor has been kind enough to spare five minutes of class time in order to allow me to introduce myself. I am currently completing my PhD at the University of Delaware studying how individuals make decisions about whether or not to support a public policy or program based on how that information is presented. This research is important because many funding agencies which support policy research and academic journal which publish policy research are seeing a rise in certain statistical methods over others. The research community needs to understand if certain statistical methods might cause individuals to support a public policy or program despite the information itself being identical. In other words, the only difference is the statistical framework used to present this information.

All that I am asking from you is your participation in completing an online survey which will not take more than 10 minutes. During the survey you will read a short vignette about research conducted for a new algebra curriculum and then answer a question about whether or not you would endorse the program. Your participation will help us understand how individuals reason about statistical information and how such information should be presented. You will also receive a \$10 Amazon gift card and be entered into a raffle to receive a \$35 Amazon gift card. If you are interested, please put your contact information (name and email address) on this sheet and expect to receive an email within the next 10 days containing a link to the survey.

## Appendix D

### DATA COLLECTION TIMELINE

<b>Activity</b>	<b>Anticipated Date</b>
Receive IRB Approval	August 2019
Identify Sample and Send Recruitment Emails	September 2019 thru March 2020
Survey Available on Qualtrics	September 2019 thru March 2020
Data Collection Ends	March 2020
Data Analyzed	March 2020 thru April 2020
Final Defense	May 2020

## Appendix E

### VIGNETTES

#### **Sample Vignette Frequentist: Moderate Evidence + Low Cost + Difficult Implementation**

A leading research firm conducted a rigorous evaluation of a new algebra curriculum. Please review the statements below summarizing their findings and answer the questions that follow.

Analysis of scores on a standardized algebra test revealed that the new algebra curriculum produced scores that were, on average, higher than scores of the control group, corresponding to an expected effect size of 0.28 ( $p = 0.25$ ). This result is considered an average effect relative to other education interventions.

However, because the result is not statistically significant ( $p > .05$ ), we cannot be confident that the new algebra curriculum improved student performance.

A cost analysis of the program revealed that the program is relatively inexpensive; a typical school district will need to spend approximately 2% of their curriculum budget in support of the program.

An implementation analysis revealed that the program may be somewhat difficult to implement because public support for the program is mixed.

### **Sample Vignette Bayesian: Moderate Evidence + Low Cost + Difficult Implementation**

A leading research firm conducted a rigorous evaluation of a new algebra curriculum. Please review the statements below summarizing their findings and answer the questions that follow.

Analysis of scores on a standardized algebra test revealed an 87.1% probability that the new algebra curriculum increased algebra scores compared to the control group, corresponding to an expected effect size of at least 0.28. This result is considered an average effect relative to other education interventions.

Conversely, there is a 12.9% probability that the treatment group reduced algebra scores compared to the control group.

A cost analysis of the program revealed that the program is relatively inexpensive; a typical school district will need to spend approximately 2% of their curriculum budget in support of the program.

An implementation analysis revealed that the program may be somewhat difficult to implement because public support for the program is mixed.

**Sample Vignette Frequentist: Moderate Evidence + High Cost + Easy Implementation**

A leading research firm conducted a rigorous evaluation of a new algebra curriculum. Please review the statements below summarizing their findings and answer the questions that follow.

Analysis of scores on a standardized algebra test revealed that the new algebra curriculum produced scores that were, on average, higher than scores of the control group, corresponding to an expected effect size of 0.28 ( $p = 0.25$ ). This result is considered an average effect relative to other education interventions.

However, because the result is not statistically significant ( $p > .05$ ), we cannot be confident that the new algebra curriculum improved student performance.

A cost analysis of the program revealed that the program is relatively expensive; a typical school district will need to spend approximately 44% of their curriculum budget in support of the program.

An implementation analysis revealed that the program may be somewhat easy to implement because public support for the program is strong.

**Sample Vignette Bayesian: Moderate Evidence + High Cost + Easy  
Implementation**

A leading research firm conducted a rigorous evaluation of a new algebra curriculum. Please review the statements below summarizing their findings and answer the questions that follow.

Analysis of scores on a standardized algebra test revealed an 87.1% probability that the new algebra curriculum increased algebra scores compared to the control group, corresponding to an expected effect size of at least 0.28. This result is considered an average effect relative to other education interventions.

Conversely, there is a 12.9% probability that the treatment group reduced algebra scores compared to the control group.

A cost analysis of the program revealed that the program is relatively expensive; a typical school district will need to spend approximately 44% of their curriculum budget in support of the program.

An implementation analysis revealed that the program may be somewhat easy to implement because public support for the program is strong.

## Appendix F

### VIGNETTE QUESTIONNAIRE

**When answering the following questions, you should be aware that bipartisan support for this algebra program has already been established. Additionally, appropriation dollars have been allocated for all school districts implementing the new algebra curriculum meaning districts will not have to spend more than what was presented in the scenario you just read.**

	Strongly agree (7)	Agree (6)	Somewhat agree (5)	Neither agree nor disagree (4)	Somewhat disagree (3)	Disagree (2)	Strongly disagree (1)
The results from this study are informative (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The results from this study are easy to understand (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Based on the results of this study, school officials are justified in endorsing this program for their school district (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Based on the results of this study, members of local school boards are justified in endorsing this program for their school district (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement (5)



Based on the results of this study, Congressional Members are justified in casting a vote, which provides financial support for the program (6)



**I. Demographic Questions for Legislative Aides:**

Q35 Please select your age

21-25 (1)

26-29 (2)

30-34 (3)

35-39 (4)

40+ (5)

---

Q36 What was your undergraduate major in college?

---



Q37 How many courses in statistics have you completed?

- None (0)
  - 1 course (1)
  - 2 courses (2)
  - 3 courses (3)
  - 4 courses (4)
  - 5 or more courses (5)
- 



Q38 Please select up to three policy areas in your portfolio of work

- Education (1)
  - Healthcare (2)
  - Labor/Employment (3)
  - Housing/Transportation (4)
  - Defense (5)
-

Q39 Are you currently employed by a Member of the House of Representatives or the Senate?

US House of Representatives (1)

US Senate (2)

Neither (3)



Q40 Please enter the email address where you want to receive your gift card

---

**II. Demographic Questions for Political Science Undergraduates:**

Q35 What is your current year at Delaware?

- First year (1)
  - Second year (2)
  - Third year (3)
  - Fourth year (4)
  - Fifth year or greater (5)
- 

Q36 What is your undergraduate major?

---

---



Q37 How many courses in statistics have you completed?

- None (0)
  - 1 course (1)
  - 2 courses (2)
  - 3 courses (3)
  - 4 courses (4)
  - 5 or more courses (5)
- 



Q38 Please select the top three policy areas of most interest to you

- Education (1)
  - Healthcare (2)
  - Labor/Employment (3)
  - Housing/Transportation (4)
  - Defense (5)
- 



Q40 Please enter the email address where you want to receive your gift card

---

## **Appendix G**

### **FREQUENTIST LOGISTIC REGRESSION MODELS**

Six separate frequentist repeated-measures logistic regression models were conducted for both samples of Congressional staffers and University of Delaware undergraduates (i.e., one model for each of the six key survey response items). IBM Corp (2018) SPSS version 26.0, a computer software for conducting statistical analyses was used to conduct the frequentist analysis. The dependent variable was the item level response and a single factor for paradigm was modeled. The exponentiated coefficient, standard deviations, and significance values are reported for each model. Frequentist repeated-measures logistic regression models were employed because actual decision making ultimately requires a yes or no answer. Hence, it is useful to consider the interpretation of this experiment from the framework of a yes or no response. To appropriately capture this dichotomization, responses from one thru four were coded zero and responses from five thru seven were coded a one.

#### **Congressional Staffers**

Table 17 presents the exponentiated coefficient, standard deviations, and significance values for each of the items for the paradigm factor. Three models revealed a statistically significant effect for the paradigm factor showing respondents are more likely to endorse a yes vote for the item if the vignette that preceded the item response was presented using the Bayesian paradigm. For all three statistically significant models, the exponentiated coefficient was in order of magnitude greater than threefold demonstrating individuals are more than three times as likely to provide a yes decision when results are presented in the Bayesian paradigm. Despite three models not achieving statistical significance, the exponentiated coefficient for these

models was in order of magnitude greater than twofold demonstrating individuals are more than two times as likely to provide a yes decision when results are presented in the Bayesian paradigm. The main effect for scenario and the interaction of scenario and paradigm factors were not tested in any of the models.

For the first item (i.e., “The results from this study are informative.”), a statistically significant effect of paradigm was found, showing a yes decision was more than three times as likely when the presented paradigm was Bayesian ( $p = 0.003$ ) and exponentiated coefficient ( $exp\ coeff = 3.41$ ). For the second item, (i.e., “The results from this study are easy to understand.”), a statistically significant effect of paradigm was not observed, with a yes decision more than two times as likely when the presented paradigm was Bayesian ( $p = 0.068$ ) and exponentiated coefficient ( $exp\ coeff = 2.17$ ). For the third item, (i.e., “Based on the results of this study, school official are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a yes decision was more than three times as likely when the presented paradigm was Bayesian ( $p = 0.029$ ) and exponentiated coefficient ( $exp\ coeff = 3.30$ ). For the fourth item, (i.e., “Based on the results of this study, members of local school boards are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a yes decision was more than four times as likely when the presented paradigm was Bayesian ( $p = 0.011$ ) and exponentiated coefficient ( $exp\ coeff = 4.55$ ). For the fifth item, (i.e., “Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement.”), a statistically significant effect of paradigm was not observed, with a yes decision more than two times as likely when the presented paradigm was

Bayesian ( $p = 0.073$ ) and exponentiated coefficient ( $exp\ coefficient = 2.19$ ). For the final item, (i.e., “Based on the results of this study, Congressional Members are justified in casting a vote, which proves financial support for the program.”), a statistically significant effect of paradigm was not observed, with a yes decision more than two times as likely when the presented paradigm was Bayesian ( $p = 0.063$ ) and exponentiated coefficient ( $exp\ coefficient = 2.31$ ).

Table 17 Overall Frequentist Repeated Measures Logistic Regression Values for Each of the Items by Paradigm for the Congressional Staffers

Variable	Exponentiated Coefficient	SE	p
Item 1	3.41	.44	.007
Item 2	2.17	.42	.068
Item 3	3.30	.54	.029
Item 4	4.55	.57	.011
Item 5	2.19	.43	.073
Item 6	2.31	.44	.063

### University of Delaware Undergraduates

Table 18 presents the exponentiated coefficient, standard deviations, and significance values for each of the items for the paradigm factor. Three models revealed a statistically significant effect for the paradigm factor showing respondents are more likely to endorse a yes vote for the item if the vignette that preceded the item response was presented using the Bayesian paradigm. For all three statistically significant models, the exponentiated coefficient was in order of magnitude greater than two and a half demonstrating individuals are more than two and a half times as likely to provide a yes decision when results are presented in the Bayesian paradigm.

Despite three models not achieving statistical significance, the exponentiated coefficient for these models was in order of magnitude greater than one demonstrating individuals are more than one times as likely to provide a yes decision when results are presented in the Bayesian paradigm. The main effect for scenario and the interaction of scenario and paradigm factors were not tested in any of the models.

For the first item (i.e., “The results from this study are informative.”), a statistically significant effect of paradigm was not observed, with a yes decision more than two times as likely when the presented paradigm was Bayesian ( $p = 0.189$ ) and exponentiated coefficient ( $exp\ coeff = 2.54$ ). For the second item (i.e., “The results from this study are easy to understand.”), a statistically significant effect of paradigm was not observed, with a yes decision more than one and a half times as likely when the presented paradigm was Bayesian ( $p = 0.510$ ) and exponentiated coefficient ( $exp\ coeff = 1.57$ ). For the third item (i.e., “Based on the results of this study, school official are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a yes decision was more than five times as likely when the presented paradigm was Bayesian ( $p = 0.003$ ) and exponentiated coefficient ( $exp\ coeff = 5.59$ ). For the fourth item (i.e., “Based on the results of this study, members of local school boards are justified in endorsing this program for their school district.”), a statistically significant effect of paradigm was found, showing a yes decision was more than two and a half times as likely when the presented paradigm was Bayesian ( $p = 0.018$ ) and exponentiated coefficient ( $exp\ coeff = 2.69$ ). For the fifth item (i.e., “Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement.”), a statistically significant effect of paradigm was not

observed, with a yes decision more than one times as likely when the presented paradigm was Bayesian ( $p = 0.510$ ) and exponentiated coefficient ( $exp\ coefficient = 1.30$ ). For the final item (i.e., “Based on the results of this study, Congressional Members are justified in casting a vote, which proves financial support for the program.”), a statistically significant effect of paradigm was found, showing a yes decision was more than three times as likely when the presented paradigm was Bayesian ( $p = 0.018$ ) and exponentiated coefficient ( $exp\ coefficient = 3.26$ ).

Table 18 Overall Frequentist Repeated Measures Logistic Regression Values for Each of the Items by Paradigm for the Undergraduates

Variable	Exponentiated Coefficient	<i>SE</i>	<i>p</i>
Item 1	2.54	.70	.189
Item 2	1.57	.68	.510
Item 3	5.59	.56	.003
Item 4	2.69	.40	.018
Item 5	1.30	.40	.510
Item 6	3.26	.49	.018

## **Appendix H**

### **CUMULATIVE RESPONSES**

Appendix H presents bar graphs showing the cumulative counts of responses for each of the six items. The cumulative counts are presented for each level of the scaled response. Visual inspection confirms the previously presented analytic results that individual respondents across both Congressional staffers and undergraduate samples generally provide a more agreeable response when the Bayesian condition is presented.

## Congressional Staffer Graphs

Figure 3 Cumulative Counts for Question 1

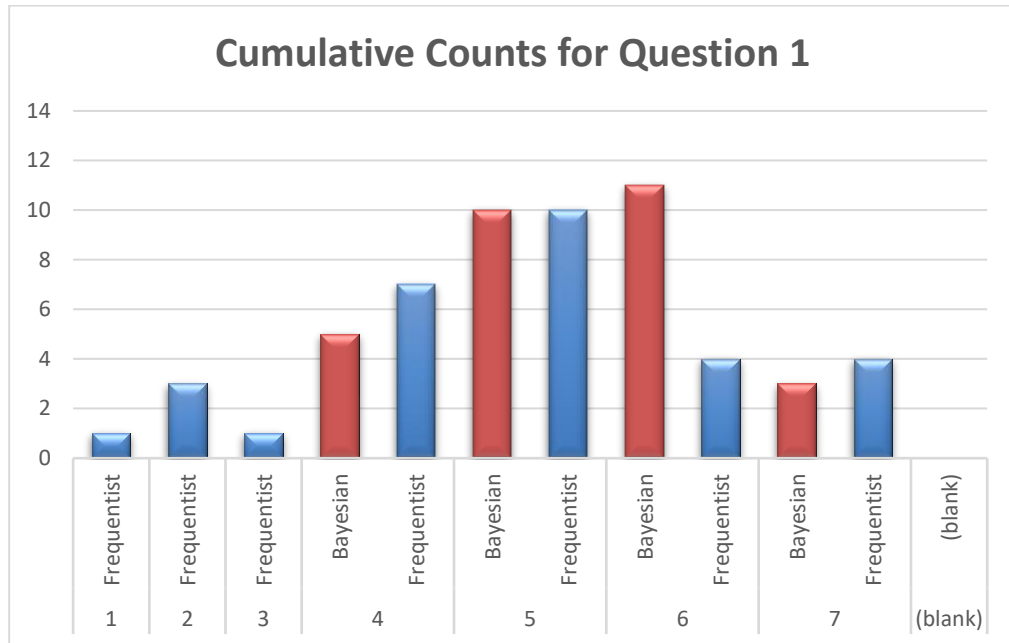


Figure 4 Cumulative Counts for Question 2

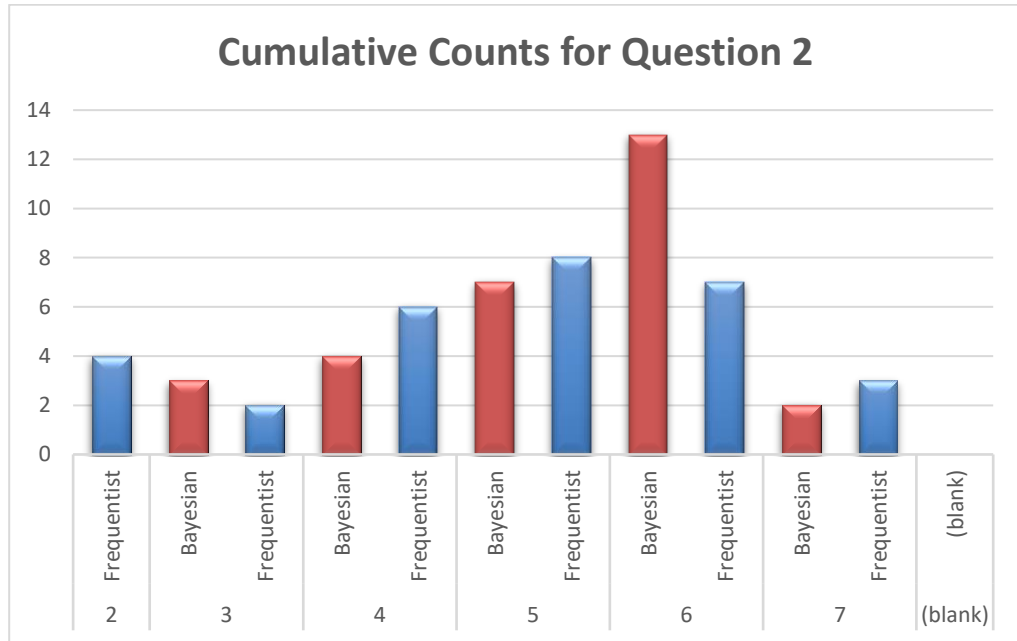


Figure 5 Cumulative Counts for Question 3

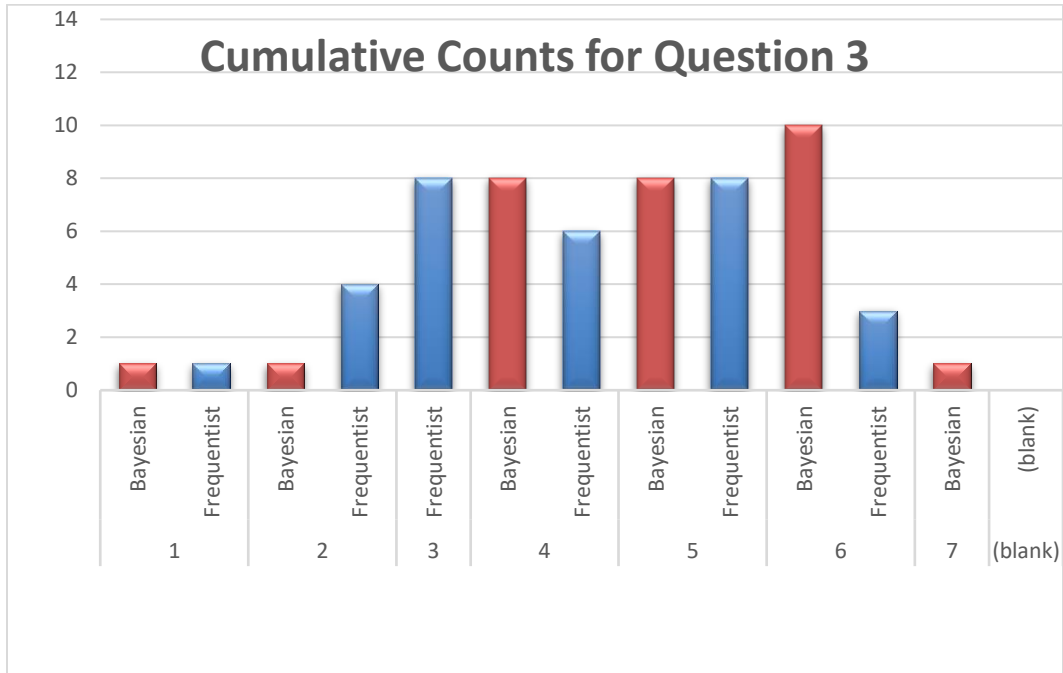


Figure 6 Cumulative Counts for Question 4

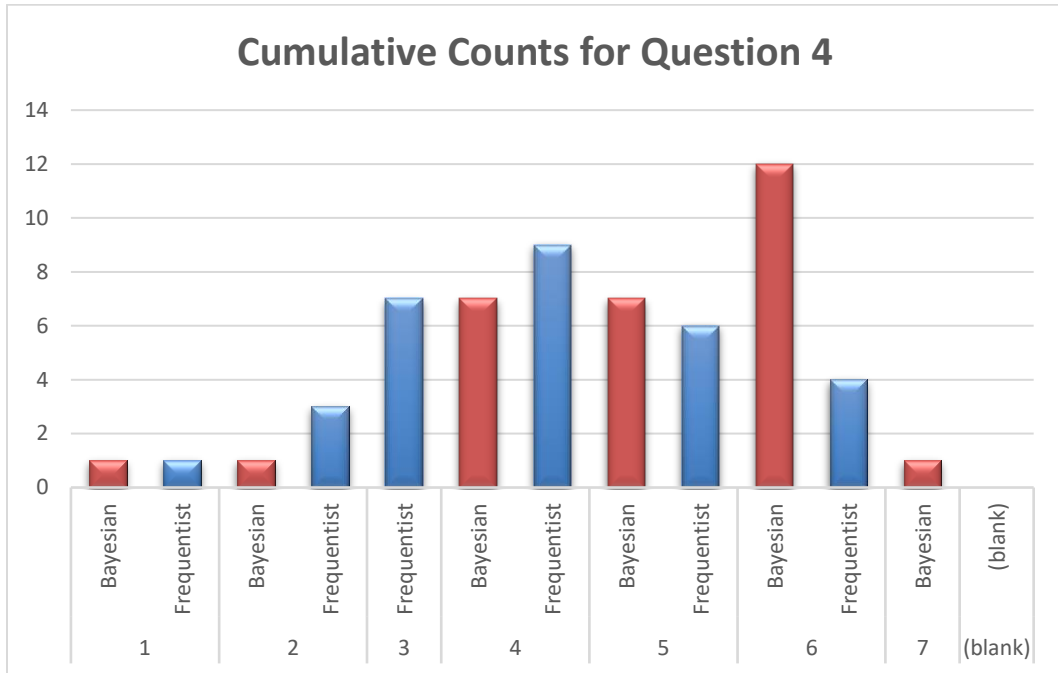


Figure 7 Cumulative Counts for Question 5

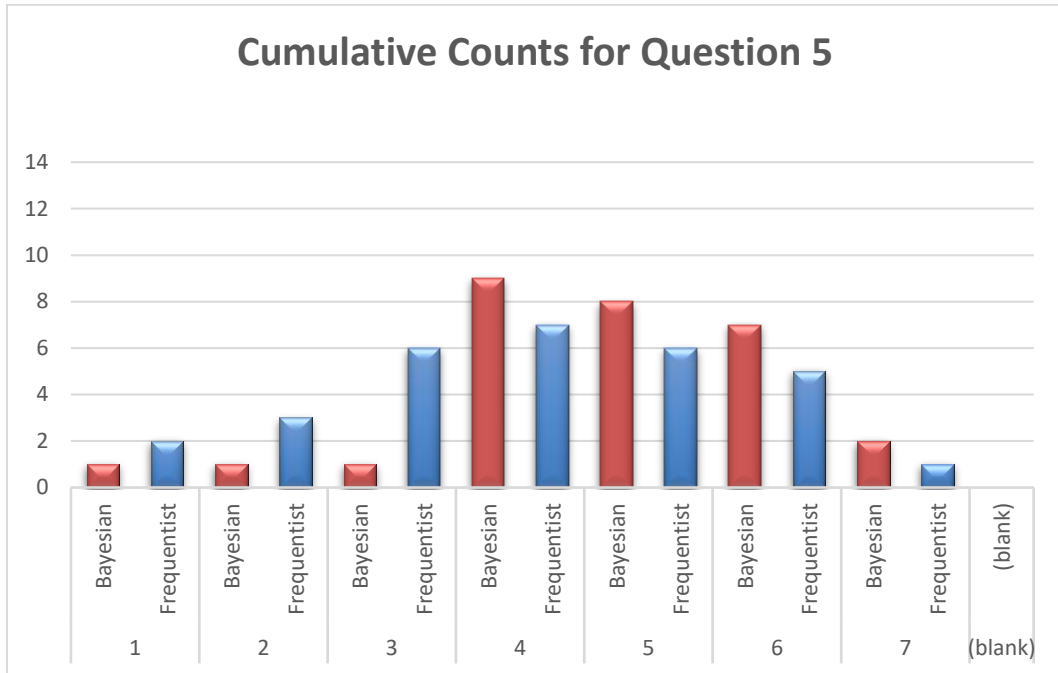
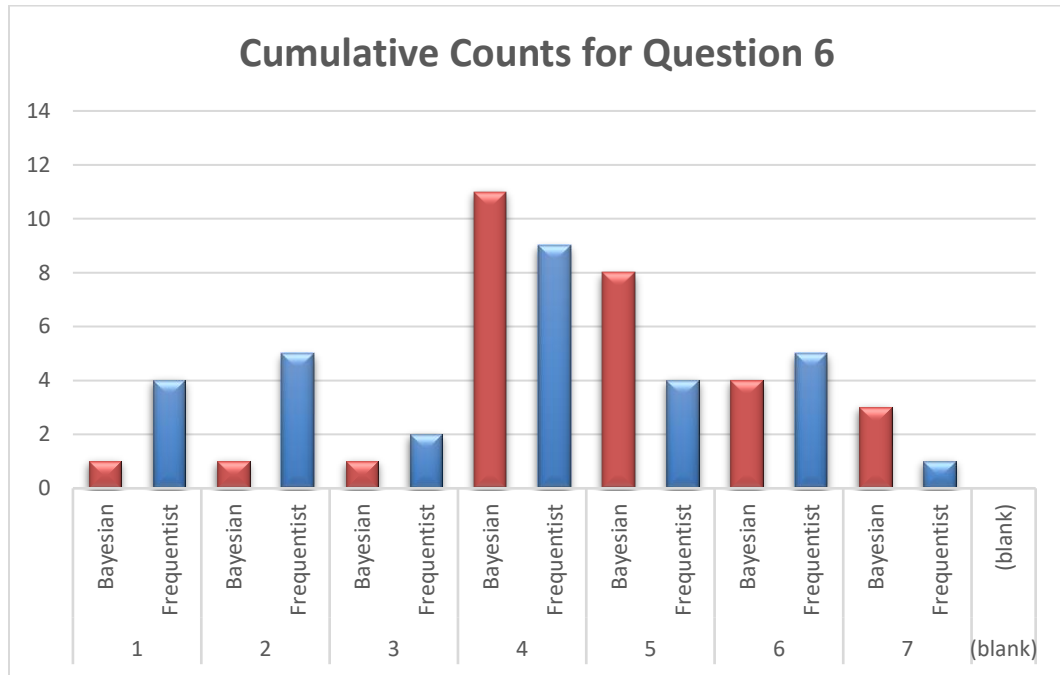


Figure 8 Cumulative Counts for Question 6



## Undergraduate Graphs

Figure 9 Cumulative Counts for Question 1

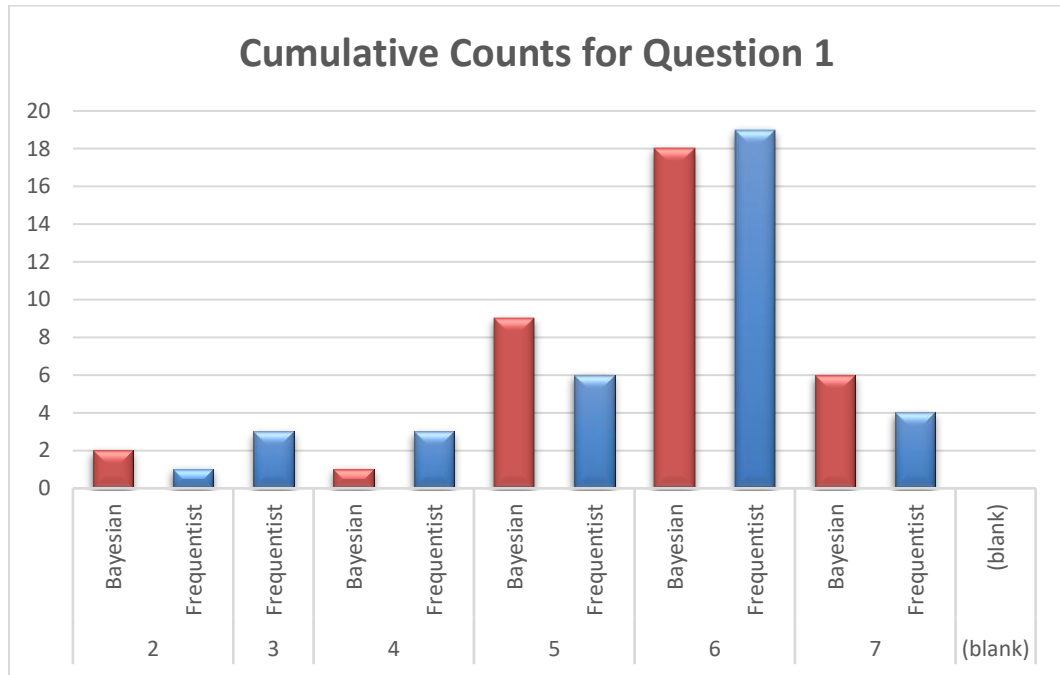


Figure 10 Cumulative Counts for Question 2

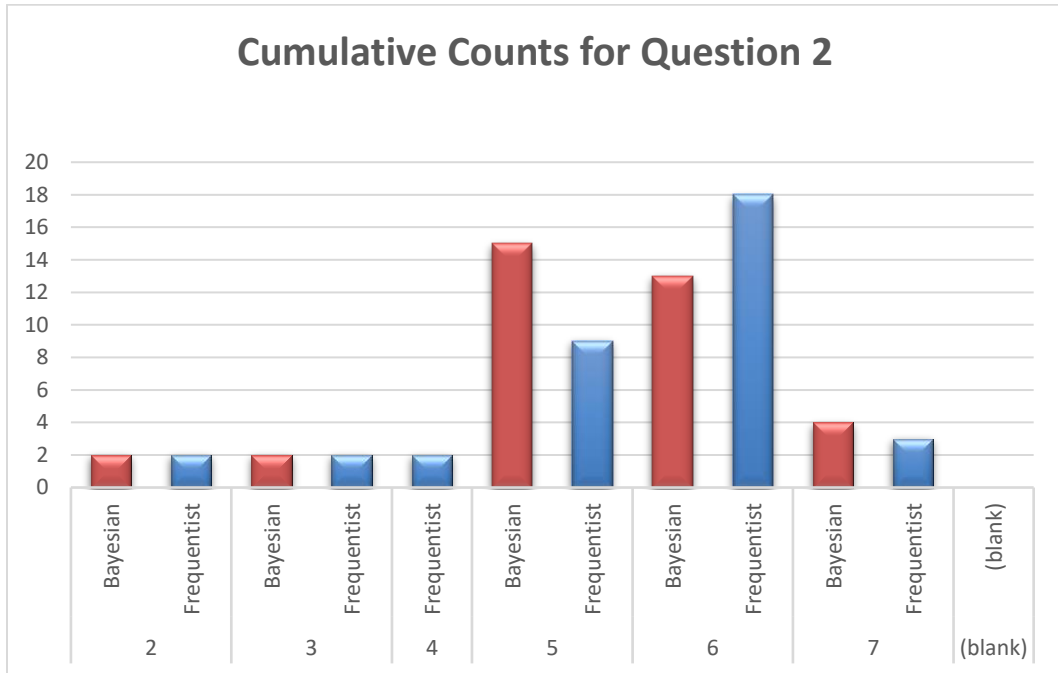


Figure 11 Cumulative Counts for Question 3

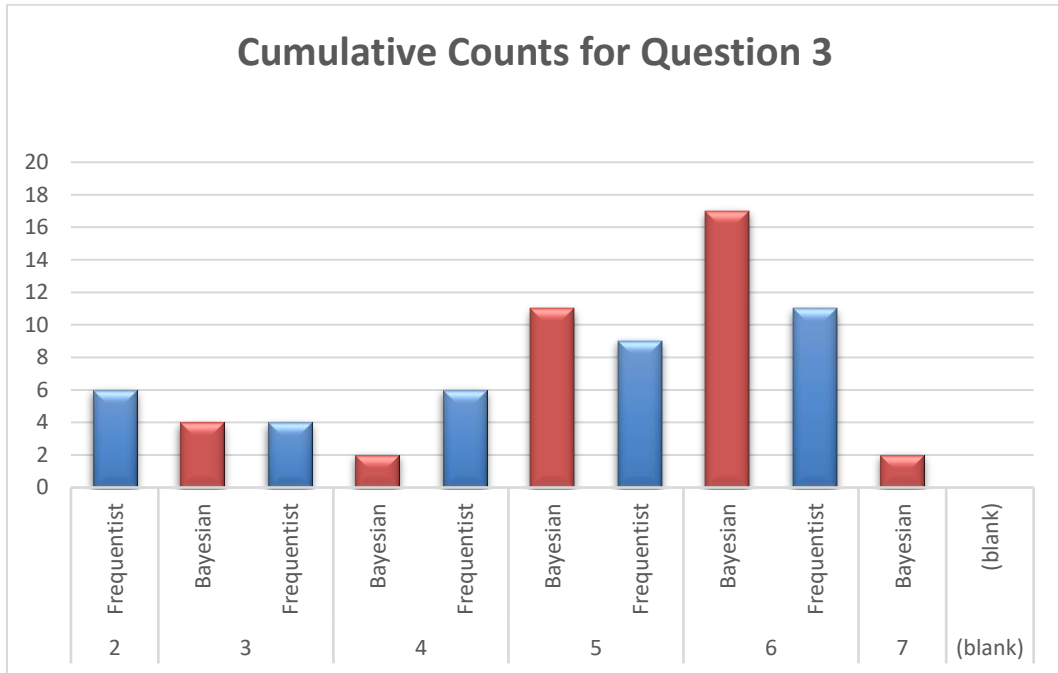


Figure 12 Cumulative Counts for Question 4

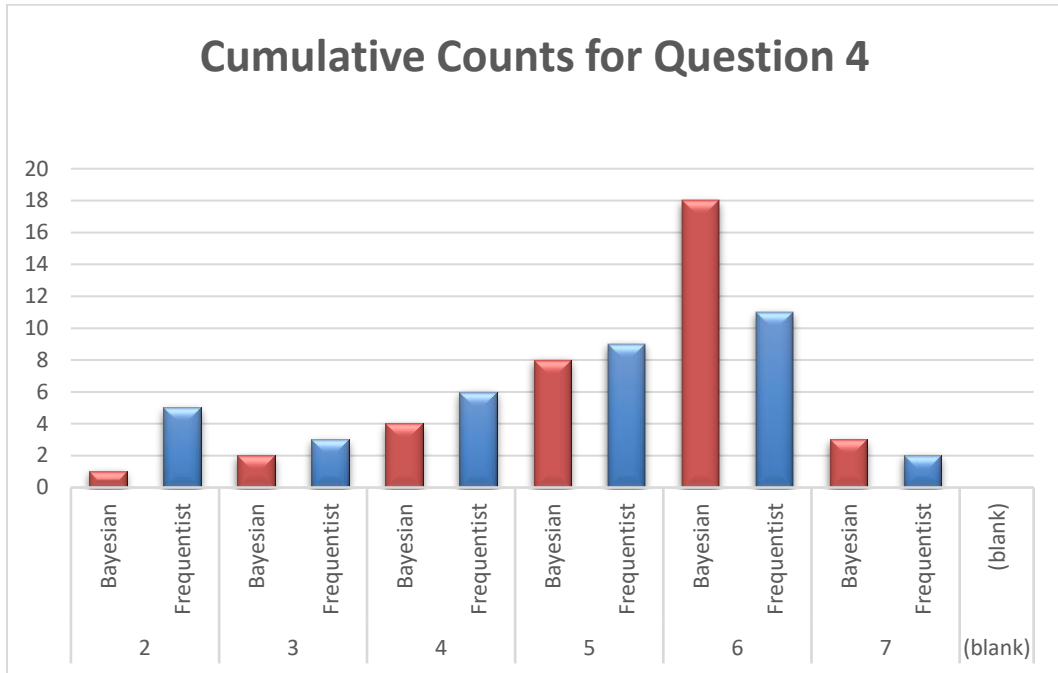


Figure 13 Cumulative Counts for Question 5

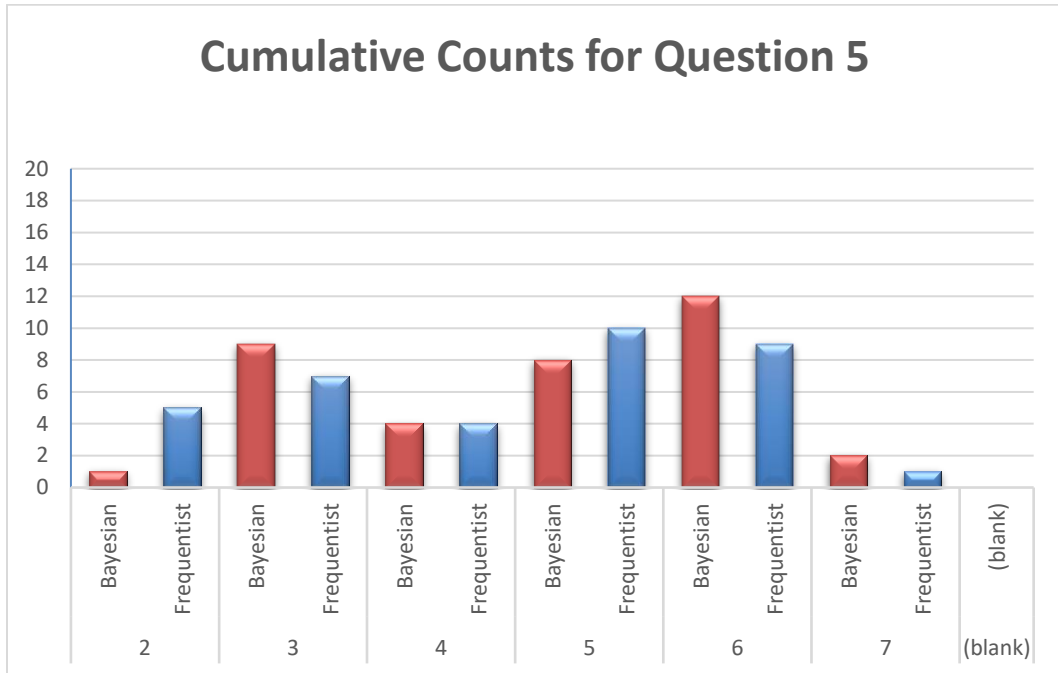
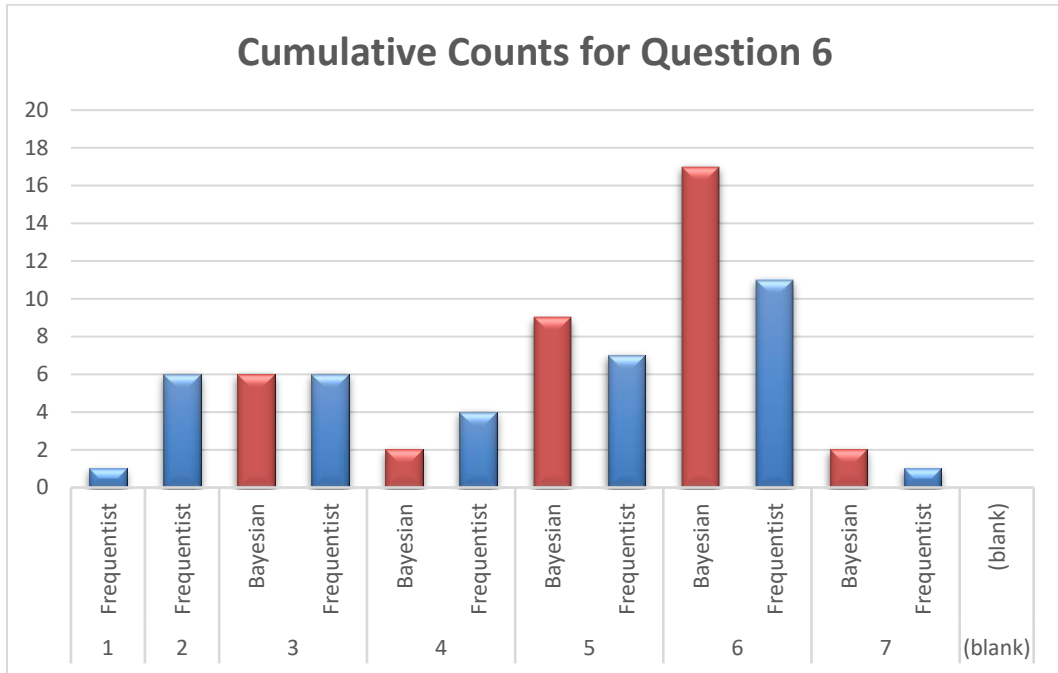


Figure 14 Cumulative Counts for Question 6



## **Appendix I**

### **FREQUENTIST ANOVA MODELS FOR POOLED SAMPLE**

Six separate two-factor repeated-measures ANOVA frequentist models were conducted on a pooled sample of both the Congressional staffers and University of Delaware undergraduates (i.e., one model for each of the six key survey response items). IBM Corp (2018) SPSS version 26.0, a computer software for conducting statistical analyses was used to conduct the frequentist analysis. The dependent variable was the item level response and the factors tested included scenario (high cost/difficult to implement and low cost/easy to implement), paradigm (Bayesian versus frequentist), group (Congressional staffer versus undergraduate) and an interaction for paradigm and group. Mean values, standard deviations, ANOVA f-values, and effect sizes are reported for the observed significant factor of group. Significant results were not observed for the factors of scenario or the interaction factor for group and paradigm. Reported values for the factor paradigm are discussed earlier in this dissertation and therefore not revisited here.

#### **Pooled Sample of Congressional Staffers and Undergraduates**

Table 19 presents the means, standard deviations, ANOVA f-values, and effect sizes for each of the items for the group factor. Only one model revealed a statistically significant effect for the group factor showing undergraduates are more likely to endorse a favorable view for the item when compared to congressional staffers. Inspection of the means demonstrates that in general, undergraduates provide a more agreeable response when compared to Congressional staffers; however, as mentioned previously, only one model achieved statistical significance for this comparison. For the first item (i.e., “The results from this study are informative.”), no statistically

significant effect of group was observed. For the second item, (i.e., “The results from this study are easy to understand.”), no statistically significant effect of group was observed. For the third item, (i.e., “Based on the results of this study, school officials are justified in endorsing this program for their school district.”), no statistically significant effect of group was observed. For the fourth item, (i.e., “Based on the results of this study, members of local school boards are justified in endorsing this program for their school district.”), a statistically significant effect of group was found, showing a more agreeable rating when the item was rated by undergraduates compared to Congressional staffers ( $p = 0.036$ ) and effect size ( $d_{rm} = +0.51$ ). For the fifth item, (i.e., “Based on the results of this study, Congressional Members are justified in lending their non-financial support for the program through writing a letter of endorsement.”), no statistically significant effect of group was observed. For the final item, (i.e., “Based on the results of this study, Congressional Members are justified in casting a vote, which proves financial support for the program.”), no statistically significant effect of group was observed.

Table 19 Overall Frequentist ANOVA Values for Each of the Items by Group

Variable	Staffer		Undergrad		F	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Item 1	5.03	1.32	5.53	1.18	3.60	-----
Item 2	4.97	1.35	5.32	1.22	1.56	-----
Item 3	4.36	1.41	4.86	1.34	3.67	-----
Item 4	4.46	1.40	5.01	1.36	4.60*	+0.51
Item 5	4.39	1.49	4.57	1.43	0.39	-----
Item 6	4.20	1.62	4.75	1.49	3.24	-----

Note.  $d_{rm}$  is Cohen's  $d$  for repeated measures. \*\*\* $p < 0.001$  \*\* $p < 0.01$  \* $p < 0.05$

## Appendix J

### BENJAMINI-HOCHBERG PROCEDURE

The Benjamini-Hochberg procedure was used to decrease the false discovery rate or the risk of type I errors (Glen, 2015). A false discovery rate of five percent was selected. Results from the adjustment procedure are provided below.

Table 20 Benjamini-Hochberg Procedure for Congressional Staffers

Variable	$p$	Rank	( $l/m$ )Q
Item 3	0.001	1	0.008
Item 4	0.001	2	0.017
Item 6	0.002	3	0.025
Item 5	0.003	4	0.033
Item 1	0.003	5	0.042
Item 2	0.015	6	0.050

Table 21 Benjamini-Hochberg Procedure for Undergraduates

Variable	$p$	Rank	( $l/m$ )Q
Item 3	0.002	1	0.008
Item 4	0.013	2	0.017
Item 6	0.01	3	0.025
Item 5	0.274	4	0.033
Item 1	0.364	5	0.042
Item 2	0.944	6	0.050

# Appendix K

## IRB APPROVAL LETTER



Institutional Review Board  
210H Hollihen Hall  
Newark, DE 19716  
Phone: 302-831-2137  
Fax: 302-831-2828

DATE: August 15, 2019

TO: Andrew Hurwitz, M.A.  
FROM: University of Delaware IRB

STUDY TITLE: [1430832-3] Is the glass half empty or is the glass half full? Understanding the role of Bayesian and frequentist statistics for influencing legislative aides' willingness to endorse programs and policies: A statistical vignette experiment

SUBMISSION TYPE: Amendment/Modification

ACTION: DETERMINATION OF EXEMPT STATUS  
EFFECTIVE DATE: August 15, 2019

REVIEW CATEGORY: Exemption category # (2)

Thank you for your Amendment/Modification submission to the University of Delaware Institutional Review Board (UD IRB). According to the pertinent regulations, the UD IRB has determined this project is EXEMPT from most federal policy requirements for the protection of human subjects. The privacy of subjects and the confidentiality of participants must be safeguarded as prescribed in the reviewed protocol form.

This exempt determination is valid for the research study as described by the documents in this submission. Proposed revisions to previously approved procedures and documents that may affect this exempt determination must be reviewed and approved by this office prior to initiation. The UD amendment form must be used to request the review of changes that may substantially change the study design or data collected.

Unanticipated problems and serious adverse events involving risk to participants must be reported to this office in a timely fashion according with the UD requirements for reportable events.

A copy of this correspondence will be kept on file by our office. If you have any questions, please contact the UD IRB Office at (302) 831-2137 or via email at [hsrb-research@udel.edu](mailto:hsrb-research@udel.edu). Please include the study title and reference number in all correspondence with this office.

**INSTITUTIONAL REVIEW BOARD**

[www.udel.edu](http://www.udel.edu)