Spring 5-14-2021

Bayesian Methods in Operational Testing: Enhancing Testing Through Combining Information

Victoria R C Sieck

Victoria R. C. Sieck

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Fletcher Christensen, Chairperson

Gabriel Huerta

Raymond Hill

Laura Freeman

Bayesian Methods in Operational Testing: Enhancing Testing Through Combining Information

by

Victoria R. C. Sieck

B.S., Mathematics, St. Edward's University, 2012
M.S., Statistics, Texas A&M University, 2018

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2021

DISCLAIMER

The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense or of the United States Air Force.

Dedication

Dedicated to my husband, without whom I could not have done this. He was my "Editor-in-Chief", my biggest cheerleader, my unwavering rock of support, and made sure I stayed awake in the long hours of the night waiting for my code to finish. His love and support made this possible!

Acknowledgments

I would like to express my deepest gratitude to my research advisor, Dr. Fletcher Christensen for taking a chance on me, supporting me, and his guidance. I would also like to thank my committee: Dr. Ray Hill, Dr. Laura Freeman, and Dr. Gabriel Huerta. Dr. Huerta provided insights into how this work might apply to other fields. Dr. Freeman began providing insights from the beginning stages. Dr. Hill not only provided his testing perspective, but supported me throughout this process with letters of recommendation and logistical support. Thank you everyone!

I would like to thank those in academia and at the national labs for their support and guidance. I am very grateful to Dr. Ron Christensen for his advice on various points presented in this work. Thank you to Ms. Ana Lombard, who provided her trucking experience to make a realistic simulated example. I would also like to thank Dr. Alyson Wilson, Dr. Rebecca Medlin, Dr. Rich Warr, Dr. Mike Hamada, Dr. Brian Weaver, and Dr. Scott Berry for taking their time to mentor me, provide their expert opinion, suggest directions of research, and overall support me in my journey.

I would also like to thank those in the Air Force who have supported my journey. Col T2 Timmerman, USAF (Ret.), who provided operational insights and support; Dr. Tom Spencer, who was pivotal in ensuring critical logistics; Lt Col Brian Stone, who provided feedback; and Mr. Dan Telford, who provided the seed for this work. Thank you to all the personnel at AFOTEC who believed in me and supported me in this endeavor. Thank you to AFIT/ENC for selecting me for a faculty pipeline position, and giving the me the opportunity to purse this research. I'm so grateful for all the support and advice I've been given in this endeavor!

Finally, I would like to thank my parents and my family for their unwavering support and love during this process. Thank you for all you've done to help me!

Portions of this work, conducted under the guidance of Dr. F Christensen, have appeared in the peer-reviewed Journal *Quality and Reliability Engineering International*. These portions are reprinted here with permission from the publisher.

Bayesian Methods in Operational Testing: Enhancing Testing Through Combining Information

by

Victoria R. C. Sieck

B.S., Mathematics, St. Edward's University, 2012
M.S., Statistics, Texas A&M University, 2018

Ph.D., Statistics, University of New Mexico, 2021

Abstract

When developing a system, considering system performance from a user perspective can be done through operational testing—assessing the ability of representative users to accomplish tasks with the system in operationally representative environments. This critical process can be expensive and time-consuming. We show how to leverage an existing design of experiments (DOE) process to construct a Bayesian adaptive design. This method allows for interim analyses using predictive probabilities to stop testing early for success or futility. Furthermore, operational environments are directly used in product evaluation. Representative simulations demonstrate reductions in necessary test events. Next, priors are built using developmental testing data. The novel proposal for creating priors using developmental testing data allows for more flexibility than the current process and demonstrates it is possible to get more precise parameter estimates. The methods presented will allow future testing to be conducted in less time and at less expense, on average.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Deliver performance at the speed of relevance.* Success no longer goes to the country that develops a new technology first, but rather to the one that better integrates it and adapts its way of fighting. Current processes are not responsive to need; the Department [of Defense] is over-optimized for exceptional performance at the expense of providing timely decisions, policies, and capabilities to the warfighter.
> - Jim Mattis, *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge* (2018)

In order to maintain a competitive edge, it is imperative that the warfighter obtain new systems in a timely manner. Without efforts to "modernize our military to make it fit for our time, we will rapidly lose our military advantage, resulting in a Joint Force that has legacy systems irrelevant to the defense of our people" (Mattis 2018). Our goal in this dissertation is to leverage existing information, such as observations seen during system testing or seen in previous system testing, to make quicker decisions about whether a new system would provide needed capabili-

ties to the warfighter. We propose accomplishing this through interim analysis and informative priors, in an effort to make those decisions effectively and efficiently.

The U.S. DoD procures these new systems through through a highly regulated acquisition process established and overseen by the U.S. Congress (*Test and Evaluation Management Guide* 2005). The DoD's acquisition process requires thorough testing of a system to ensure that engineering specifications are met for the system and that users can perform their intended mission with the system, ultimately evaluating whether a system is mission capable or not before employing it in its expected operational environment. These technical specifications and performance attributes are referred to as *requirements*, which testers use to define response variables of interest to evaluate (Joint Chiefs of Staff 2018). This formal test and evaluation process ensures the DoD procures systems that the warfighters needs and can use. While this test and evaluation process has many parts, we will consider two overarching phases: developmental testing (DT) and operational testing (OT).

The first governmental testing of a system is DT. The primary goal of DT is to determine if a system can obtain established technical requirements. Generally evaluating distinct characteristics of the system itself, DT "assesses *if* and *how* the system works" (National Research Council 1998). To accomplish this, DT is largely conducted in laboratories or in controlled environments, and experts are the system operators (*Test and Evaluation Management Guide* 2005). Early in DT a system is often still undergoing development, in part to identify problems early; therefore, testing of components or sub-systems may occur. As DT continues, and the system under test matures, prototypes (or even production representative systems) may be evaluated (*Test and Evaluation Management Guide* 2005). Information and insights gained during this testing can drive system fixes and upgrades, resulting in a system that evolves throughout the course of DT.

After DT establishes that a system meets DoD's technical specifications, the sys-

tem under test proceeds to OT. In this phase, the focus is on mission accomplishment in the system's anticipated operational environment; operational testers evaluate a production representative system to determine if users can accomplish their intended missions while employing the system (Kendall 2015). OT "assesses *when* and *where* the system will work" (National Research Council 1998). Given this difference in focus, the requirements that are evaluated in OT are usually different from the requirements evaluated in DT; therefore, it is possible for response variables of interest in DT to become latent variables or no longer be of interest in OT. OT seeks to evaluate a system's performance under a multitude of diverse operational conditions (e.g., weather, time of day), so as to assesses the impact of those conditions on mission accomplishment. This is achieved by testing in operationally realistic environments with operationally representative users of the system. Unlike the system experts used in DT, the system operators in OT have only had training on using the specific system that is representative of the training personnel expected to employ the system in future real-world operations would have (*Test and Evaluation Management Guide* 2005). This difference, in addition to the complexity of an uncontrolled operational environment, can increase the variability seen in system performance within OT relative to DT. Ultimately, OT attempts to replicate real-world operations to the maximum extent possible to ensure decision makers have operationally representative information about a system to make determinations about procurement.

To better illustrate DT and OT, consider a hypothetical trucking company that wants to procure a cellphone for their drivers to use as a business phone on the road. DT might begin with evaluating required engineering specifications for a component: for instance, testing might occur on the processor to ensure that the material used meets durability and thickness specifications. As DT progresses, testers might use a cellphone prototype to evaluate how long the battery lasts when the phone was turned on. If the battery life did not meet specifications, the phone could be re-engineered to include a new battery. Perhaps later still, an expert might test the

camera-memory interface to ensure the phone can store pictures in the memory and be accessed later. The phone would also be tested to determine if it can operate with other external systems. One such example would be to ascertain if the phone could sync to a truck via a bluetooth connection or a cable connection.

After the conclusion of DT, the cellphone would proceed to OT to evaluate the phone's ability to support mission accomplishment when employed by real-world users. To facilitate this, OT might evaluate operational requirements by giving the phone to a truck driver to operate. One evaluation could be to determine if a driver could use a map application while driving in conjunction with a making a hands-free phone call. Another may be attempting to download a map, and then use it in a location with no service. Other requirements would include how helpful customer service was in trouble-shooting an issue that arose with the phone, or how long the battery lasted when 12 internet tabs were open and applications were running in the background while in and out of roaming service. Furthermore, a determination could be made regarding how resilient the phone was to a driver tossing it into a bag or dropping it on the ground accidentally throughout the day.

Similar to a phone that will not provide connectivity in a no-coverage area but can still accomplish daily needs, OT must carefully consider if the user can still accomplish a mission at an adequate level, even should all requirements not be fully met (especially if those requirements are not obtainable, such as in a no-coverage area). This is an assessment that OT is uniquely able to provide, supplying decision makers with critical information about a system. Conducting an operational test to obtain this operationally representative information is resource intensive; however, it is an expense that is justified, given the impact of the data (National Research Council 1998). Without these insights, any problems with a system may not be found until it is employed in the operational environment—potentially endangering both lives and mission accomplishment.

In 1998, the National Research Council found that the use of statistics in test and evaluation "differ[ed] substantially from best practice, to the detriment of effective operational test and evaluation" and concluded that state-of-the-art statistical methodologies would enhance testing within the DoD. In 2010, the Director of Operational Test and Evaluation in the Office of the Secretary of Defense published DoD guidance which echoed the need to expand the use of statistics in test and evaluation. It detailed the need for defensible and technically adequate tests, concentrating on design of experiments (DOE) as one such statistical method to support OT (Director Operational Test and Evaluation 2010; Freeman and Warner 2018; R. T. Johnson et al. 2012). Using DOE ensures testers collect the right amount of data, in the appropriate places, to characterize a system's performance across operational conditions and make objective conclusions based on the data obtained (Montgomery 2012; Freeman and Warner 2018). DOE is also critical in creating operational tests that make best use of limited resources (R. T. Johnson et al. 2012).

With the amount of resources driven by time and cost constraints, limited resources, however, can lead to an operational test that gathers insufficient information to make a definitive conclusion with a reasonable degree of certainty (National Research Council 1998; National Research Council 1994; National Research Council 2004). To address this issue, it has been recommended that all available information be used in OT—such as information from DT (National Research Council 1998; National Research Council 2004; National Research Council 1994). However, as can be seen in the hypothetical example, there are many differences between what DT and OT evaluate. For a system proceeding through the phases of the acquisition process, these differences can result in related, but not identical (or, more generally, exchangeable), data. While the relatedness of the data implies that DT could inform OT in some manner, the potential for non-exchangeable data presents challenges—even if DT and OT are interested in the same response variable. These challenges include:

- OT is conducted in operationally realistic environments with operational users; DT need not be conducted in such a manner, and can be conducted in controlled or artificial environments using experts in a system.
- OT uses an production representative system; DT can use system components or early, less mature, versions of a system.
- OT focuses on evaluating whether a user can accomplish a mission while using a system; DT focuses on evaluating whether system specifications have been met.
- It is possible that DT, especially later in system development, can be conducted in a way that minimizes the effect of the differences created by the above challenges. However, even if DT and OT are conducted in the same environment, with the same system, and evaluate the same response variable, the strictly controlled manner in which a system is employed in DT could result in different outcomes than in OT, where operationally representative users employ the system as they see fit.

These challenges highlight differences in environment, system maturity, test focus, and system employment—differences that must be considered before using DT data to inform OT, to ensure that conclusions about a system are made using information that is representative of how a system is expected to perform in an operational environment. Recognizing that certain DT information may be beneficial in OT, the DoD's current OT framework has a process for qualifying appropriate non-OT data for use in OT. "For data to be qualified for OT, the data must have been collected using production representative equipment, with representative operational users, employing operational [tactics, techniques, and procedures (TTPs)]" (Department of the Navy 2019). While this is one step towards incorporating DT information, these restrictive conditions still prevent the use of all relevant

information that DT could provide. This can leave operational testers spending limited resources to capture OT data that might be unnecessary, which can result in allocating resources in a sub-optimal manner or in having insufficient data at the end end of testing to evaluate a system.

Bayesian methods are ideal for scenarios in which there is insufficient information available from a test, as all information thought to be relevant can be easily incorporated into analysis. By incorporating additional information into an analysis, the standard deviation for model parameter estimates can be improved, making it possible for conclusions to be made with the degree of certainty required when there is limited data (National Research Council 1994). However, prior to 2015, "no one [had] yet capitalized on the knowledge that can be gained when one properly combines information across all of these test venues" (Dickinson et al. 2015). Dickinson et al. (2015) present a case study for combining DT data into OT through Bayesian methods, leveraging consistent data collection across both test phases and commonalities within a family of similar systems, and demonstrate that it is possible to get more precise model parameter estimates than when using OT data alone. Alternatively, previous information can also be incorporated into a Bayesian analysis through prior distributions, as was done by Dewald et al. (2016). By using summary statistics from previous computer experiments to create a prior for a live test, they highlight how Bayesian methods can improve the DoD's current OT framework by allowing for the possibility of improved precision of model parameter estimates.

Combining information across test phases can address the case when OT alone is insufficient to make a definitive conclusion with a reasonable degree of certainty. By allowing information to be combined, an additional question can be considered—do testers know enough about the system from relevant, previous testing that the operational test design can be reduced? Given "the high cost of many weapon systems, and the substantial cost of testing them, even modest improvements in operational test-

ing by using the most appropriate statistical methods can lead to more efficient use of public funds and considerable improvements in reliability and effectiveness of the systems deployed" (National Research Council 1998). One statistical method that could be used to allow for more efficient use of funds is to incorporate interim analysis into OT, with the intent to stop OT early if enough information has been obtained to make decisions about system procurement. Not only is combining other information into analysis foundational to the Bayesian framework, but "Bayesian inferences are more flexible [than frequentist methods] in that they can be updated continually as data accumulate", making interim analysis a natural fit within a Bayesian analysis (D. A. Berry 1993). Finally, as will be discussed in Chapter 2, OT already uses informal methods to incorporate subject matter expert (SME) opinion into test design (National Research Council 2004); the Bayesian framework not only formalizes this process, but it extends the process to allow for all relevant information to be incorporated.

These concepts lay the foundation for the overarching proposition of this research: moving the DoD's current OT framework from a frequentist paradigm into a Bayesian paradigm, introducing adaptive testing principles into OT, and incorporating DT information into OT in order to make OT more effective and efficient. Chapter 2 begins with an overview of the current OT framework, followed by a simulated example that will be used throughout this research. The Bayesian framework is then overlaid on the OT paradigm. Furthermore, the operational environment is directly incorporated into the evaluation of a system. The chapter concludes with using the previously introduced simulated example to implement the proposed method. Chapter 2 ultimately proposes a way in which Bayesian methods can be employed the OT environment.

Chapter 3 begins with a literature review of specific work within clinical trials using predictive probability. Chapter 2 is then leveraged to develop a method for

adaptive OT that permits stopping test early, allowing for effective and efficient use of test resources. This section broadens the clinical trials work to OT and extends that work to a fully Bayesian method for a continuous response with an ANOVA structure. Chapter 3 concludes with the simulated example, demonstrating the utility of the method. The adaptive methods proposed in this chapter provide a novel approach to accomplishing OT, allowing for more efficient use of resources.

Chapter 4 begins with an in-depth discussion of priors. Having established a method for adaptive OT in the previous chapter, a novel method for using DT information to create informative priors for OT is then presented. By incorporating DT information, it is possible to get smaller standard deviations for model parameter estimates, which allows for stopping OT earlier than in previous chapters. Chapter 4 proposes a new informative prior that appropriately accounts for the differences between DT and OT, and is more suitable for use in OT than previously proposed priors.

Chapter 5 will summarize what the previous chapters accomplished and discuss areas for future research. The framework presented in this research allows operational testers to capitalize on additional statistical capabilities to further enhance the effectiveness and efficiency of OT. In addition to greater flexibility within testing, Bayesian methods more directly answers the question that OT seeks to address: will the user be able to accomplish the mission while using the system under test?

Chapter 2

Operational Testing from a Bayesian Perspective

Chapter 2 proposes transitioning the Department of Defense's (DoD) current operational testing (OT) paradigm into a Bayesian framework and introduces a method for directly incorporating the operational environment in analysis. Section 2.1 introduces the current OT paradigm and a simulated example that will be used throughout this research. Section 2.2 then shifts the OT paradigm into a Bayesian framework, proposing a new approach to measure evaluation and providing the groundwork for using adaptive testing principles in OT (Chapter 3). Section 2.3 implements the proposed process using the simulated example, and compares the results to the results from the current process in Section 2.1. A review of pertinent concepts within the Bayesian framework for the methods presented in this and later chapters can be found in Appendix A.

2.1 Department of Defense's Current Operational Testing Framework

Before shifting to a Bayesian perspective, the current process is introduced. This section provides a brief overview of the current OT process, followed by a simulated example implementing this process. In Section 2.3, results from this section will be compared against the results from the proposed method of this chapter For a more detailed discussion of requirement creation and development in a DoD context, see the Joint Chiefs of Staff's *Manual for the Operation of the Joint Capabilities Integration and Development System* (2018), here after referred to as the *JCIDS Manual*. For a more detailed discussion of the test design and system evaluation process, see the Department of the Navy's *Operational Test Director's Manual* (2019), hereafter referred to as the *OTD Manual*.

2.1.1 Creating an Operational Test and Evaluating a Measure

Recall from Chapter 1 that requirements are technical specifications and performance attributes used to evaluate a system. Requirements are foundational to the acquisition process—they define what the warfighter needs, which creates a framework for DT and OT efforts. Requirements detail, among other things, a *response variable* for a task or attribute of a system and a related *response threshold value*. Response threshold values "represent the value below which [system] performance would require reevaluation of military utility" in the defined mission area (*JCIDS Manual*). When evaluating a system, requirements are used to establish (what testers refer to as) *measures*. A single measure is the combination of a function of a response (referred to as a *parameter of interest* and represented by ϕ) and an associated response

threshold value (represented by ϕ_0).

In addition to being the foundation for a system's evaluation, requirements are also the foundation for statistical test designs within OT. Requirements that are deemed critical inform the selection of response variables for experimental designs (a single operational test program may be made up of multiple experimental designs, but this research will focus on one critical response variable and the associated experimental design to illustrate our proposed methods). After using requirements to choose a response variable, testers select *factors* and *levels* that are judged to be operationally relevant to the response (*OTD Manual*). Following the classical *design of experiments* (DOE) process, factors that are recordable are conditions that may have an effect on the response variable, but may be uncontrollable during test execution (e.g., weather). Design factors are factors that are ultimately selected to influence the test design (Montgomery 2012). Augmenting the classical DOE process, OT evaluates the selected factors and levels to understand their potential impact on a response variable during the test design process. First, the potential *effect* that changing from one level of a factor to another level might have on the response variable is examined and categorized as high, medium, or low. Next, the anticipated *likelihood of encountering* a given level in the system's operating environment is examined and quantified. This process is referred to as *factor prioritization* (*OTD Manual*).

Using the results from factor prioritization, an experimental design can be created, based on guidance in the *OTD Manual*. Test designs are constructed so all main effects and two-way interactions can be identified, with 80% power and a confidence level of 80% (*OTD Manual*). However, to calculate power, an appropriate effect size must be selected. This may be determined by subject matter experts (SME) input or through considering historical data; regardless of how it is selected, the effect size should have practical meaning for a system and its intended mission

(*OTD Manual*). After an operational test is completed and data are obtained, the most granular piece of system evaluation begins—evaluating measures.

To evaluate a measure, operational testers frequently combine information from different operational environments (i.e., different factors and levels) into a summary statistic that estimates a parameter of interest. Selecting a grand mean to estimate the parameter of interest is a common choice in OT (National Research Council 1998; Freeman and Warner 2018). By using a summary statistic, evaluating a measure is based on the entire space, as defined by the factors and levels, rather than group means. After obtaining test data, the parameter of interest, ϕ, is calculated and compared to the threshold value, ϕ_0. If $\phi \geq \phi_0$, the measure is evaluated as *met*; if $\phi < \phi_0$, the measure is evaluated as *not met*.

This research assumes that the current design process is robust enough to evaluate a system under test; furthermore, the methods proposed in this research focus on the most granular evaluation that occurs within OT—evaluating a single (critical) measure as met (i.e., concluding $\phi \geq \phi_0$) or not met (i.e., concluding $\phi \leq \phi_0$). The complete system evaluation (evaluating if the system is *mission capable*) is an area for future work, and discussed further in Chapter 5.

2.1.2 Simulated Example: Electric-Semi Truck

To illustrate the current OT process, recall the hypothetical example about the truck company who wanted to procure cellphones, illustrating the differences between DT and OT; we will continue to use this company to illustrate the concepts presented throughout this research, but through the procurement of a different system. In Section 2.3, the results from this section will be compared to the results from our proposed method. This hypothetical company's mission is to transport products, both regionally and across-country, on a pre-specified timeline. The CEO recently

announced an initiative to become a "green company," and a cornerstone of that initiative was to transition their 600+ diesel semi-trucks into a more environmentally-friendly fleet. After conducting an assessment of the infrastructure across the country, the company determined that it could support a fully electric semi-truck operation. The company intends to acquire the selected electric semi-truck by mirroring the DoD acquisition process.

This electric semi-truck example is a simulated example; by using simulated data, various distributions of the response can be considered, providing insights into how the method performs across a range of possible scenarios. The OT data generation processes for this example can be found in Appendix B; true parameter values are shown to compare with results, but are not used in the method itself. The simulated response values are based on publicly-proposed ranges for vehicles of a similar class being put forward by companies currently developing such a vehicle.

Beginning the OT process, company representatives, truck drivers, and other SMEs met to determine which established electric semi-truck requirements were appropriate for OT to evaluate. Requirements included topics such as training requirements for maintainers and truck operators, hardware requirements for connecting to charging stations, software requirements for the electric engine and dashboard, and many other requirements. This example focuses on one critical requirement in particular: "The average range of an electric semi-truck on one charge should be at least 400 miles."

Using this requirement, the response variable "mean number of miles traveled" was selected as the basis for an experimental design, and factors and levels that were operationally relevant for the response variable were selected (see Table 2.1). Operationally, wind is a factor that is classified as uncontrollable—therefore, wind is not randomized within the experimental design, but will be measured during the test. Wind is also an example of a factor that encompasses many aspects—aspects

such as direction and strength of the wind. The test team determined that binning numbers from the Beaufort Scale (Encyclopaedia Britannica 2017) was similar to the information available during an operational employment of the electric semi-truck and defined the following levels:

- Good: Beaufort Numbers 0-3 in any direction and 4-6 in tail wind direction
- Moderate: Beaufort Numbers 4-6 in head wind and cross wind direction
- Poor: Beaufort Number of 7 or higher

After determining the factors and levels, the test team accomplished factor prioritization. The potential effect of changing the level (e.g., moving from hilly to flat terrain would highly impact the response) was determined and can be found in Table 2.1. Next, the anticipated likelihood of encountering each level in Table 2.1 was evaluated. For example, for every nine routes the company had, four of their routes were in hot climates and five were in moderate climates. Similar assessments were made for all other factors and levels and can be found in Table 2.1.

Factor	Levels	Effect	Likelihood
Terrain	Hilly Flat	High	50% 50%
Temperature	Hot (>70°F) Moderate (70°-50°F)	Medium	4/9 5/9
Wind	Good Moderate Poor	Medium	1/3 1/3 1/3
Payload Type	Refrigerated Non-Refrigerated	High	50% 50%
Weight	Heavy (≥40k lbs) Light (<40k lbs)	High	50% 50%

Table 2.1: Factor Prioritization for Electric Engine

After factor prioritization, an experimental design could be selected. Using the statistical analysis software package Minitab, testers selected a 2^4 full factorial with five replicates that included main effects and two-way interactions (excluding wind) for the experimental design, resulting in 80 test events. This design has a power of 80%, with an 80% confidence level, to detect a difference of 50 miles with a standard deviation of 100 miles.

Finally, for the requirement used in this example, the test team determined one measure to be the following: ϕ is the mean number of miles traveled and $\phi_0 = 400$ miles. Comparing the resulting ϕ to ϕ_0 would determine if the system met the measure or not: if $\phi \geq \phi_0$, the test team would evaluate the measure as met; if $\phi < \phi_0$, they would evaluate it as not met and a discussion of the impact on the mission would ensue.

The test was then executed as designed, without any missing or censored data; after all the data were obtained, measures were evaluated. To highlight how the evaluation of the identified measure changes depending on different distributions of the response, 21 data sets were generated as described in Appendix B. Results for the identified measure using the current OT process are presented in Table 2.3 (Section 2.3), after all components of the proposed method have been introduced; however, for the 21 data sets generated, all 21 would evaluate the measure as met under the current OT paradigm. Therefore, all would lead to the conclusion that the electric semi-truck's average range on one charge was greater than 400 miles.

2.2 Operational Testing within a Bayesian Framework

Having introduced key concepts within the current OT process, this section details how to shift the OT paradigm into a Bayesian paradigm. As previously highlighted, this research assumes the process outlined in Section 2.1 develops an adequate test to evaluate system performance; as such, this section lays the groundwork for efficiencies that can be gained in the current process by shifting from a frequentist perspective to a Bayesian perspective.

From the Bayesian perspective, this chapter provides the ability for testers to calculate the *direct probability* of ϕ obtaining ϕ_0 (i.e. $\Pr_{\phi|X}(\phi \geq \phi_0)$; in contrast to the *indirect probabilities* in a frequentist analysis, i.e., *p-values*) (S. M. Berry et al. 2011; D. A. Berry 1993). Consider testing the hypothesis: $H_0 : \phi \geq \phi_0$ versus $H_1 : \phi < \phi_0$. Bayesian inference can provide the direct probability, $\Pr_{\phi|X}(\phi \geq \phi_0)$, after a test has been completed and some data, X, have been obtained—that is to say, the probability that ϕ is greater than or equal to ϕ_0, given X. In contrast, frequentist methods are indirect probabilities. Statistical tests, such as z-tests or t-tests, are used to calculate a p-value (or compared against a similarly computed critical value) to determine if H_0 should be rejected or should fail to be rejected. The interpretation of a p-value is the probability of obtaining a result as extreme or more extreme than the results obtained, assuming H_0 is true—a statement that can be confusing to non-statisticians, and is often (incorrectly) interpreted as the more understandable Bayesian direct probability (D. A. Berry 1993). Shifting OT into a Bayesian framework not only allows for direct probabilities to be calculated, it also provides the foundation for this research—exploring more efficient OT practices. In Chapter 3, predictive probabilities are used to consider stopping an operational test early, based on interim results. Chapter 4 incorporates developmental testing (DT)

information, which is disparate from OT information but still related. Both chapters are a byproduct of the shift to a Bayesian perspective presented in this chapter.

This section proposes augmentations and additions to the current process to support a Bayesian analysis, while still seeking to leverage the current process to the greatest extent possible. The section begins with augmenting the current OT process, followed by a method for explicitly analyzing data from an operational perspective. It concludes with how to evaluate a measure in a Bayesian analysis.

2.2.1 Augmenting the Current Test Design Process

While the current OT process does not require the construction of a statistical model to evaluate a measure (especially if a grand mean is used), the underlying structure of the test design process can be seen through the lens of an ANOVA model for the response, where each factor level (and two-way interaction) corresponds to a parameter in the model. Explicitly using an ANOVA model provides a framework for understanding factors and levels, and their impact on the response variable. Using a reference cell model, the following model form is proposed:

$$y_{\{h\}p} = \mu_{\{h\}} + \epsilon_{\{h\}p}$$

where $\{h\}$ represents the set of indices for the parameters in the model, $\mu_{\{h\}}$ is a linear combination of main effects and interactions defined by $\{h\}$, and p represents the replicate of the observation. We will use H to indicate the space of all allowable sets of indices, such that each h is an element of H. For example, consider a two-way ANOVA model (each factor having two levels) with the response variable $y_{ijp}|\mu_{ij}, \tau$; then $\mu_{ij} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{(ij)}$ for $i = 1, 2$ and $j = 1, 2$. In this example, η is the baseline parameter representing the first level of each factor. The constraints of this reference cell model are all model parameters at the first level, or with at least one factor at the first level for interactions, are equal to 0. This provides

the interpretation of α_2 as the change in the group mean from μ_{11} to μ_{21}; that is, $\mu_{21} - \mu_{11} = \alpha_2$. Assuming that the errors are *iid* and normally distributed with a mean of 0 and a variance of $\frac{1}{\tau}$ (where τ is referred to as the precision), the distribution of the response within each group (as defined by the factors and levels) can be written as

$$y_{\{h\}p}|\mu_{\{h\}}, \tau \overset{iid}{\sim} N\left(\mu_{\{h\}}, \frac{1}{\tau}\right).$$

In the Bayesian framework, each parameter in the model is considered random and unknown; therefore, all model parameters require *prior* distributions. After the completion of factor prioritization, discussions that inform and result in prior selection for the model parameters should be held—a natural extension of the current process that determines whether varying the levels of a factor potentially cause a high, medium, or low effect on the response. While Chapter 4 addresses the use of informative priors using DT information, Chapters 2 and 3 use independent reference priors. Instead of diffuse reference priors that strive to provide no information about model parameters (and can cause computational issues, as will be addressed in Chapter 4), reference priors that incorporate information about the bounds of the physical world (also known as weakly informative priors) ensure that the resulting posterior distribution is reasonable (Gelman et al. 2014). Section 2.3 will demonstrate how these priors can be developed, with the simplifying assumption that the priors on the model parameters are all mutually independent.

After defining an ANOVA model and selecting priors, a likelihood for the data to be obtained can be defined. Letting ξ represent the collection of model parameters and Y represent the set of observations, the form of the likelihood is:

$$L(\xi, \tau|Y) = \prod_{h \in H} f(y_{\{h\}p}|\xi, \tau).$$

Given the prior for the model parameters and the likelihood for the data, the posterior then has the form:

$$p(\xi, \tau|Y) \propto L(\xi, \tau|Y)\, p(\xi, \tau)$$
$$= \Big[\prod_{h \in H} f(y_{\{h\}p}|\xi, \tau) \Big] p(\xi, \tau).$$

The shift to a Bayesian perspective, as proposed in this section, has augmented rather than altered the current process; therefore, this proposal will result in selecting the same experimental design as the current test design process. While the method presented in this chapter can be used with any experimental design, this is not the case for the adaptive OT method described in Chapter 3. Experimental designs that can benefit from both methods include fully randomized designs (e.g., factorial designs and optimal designs) and designs that block on replicates.

2.2.2 Analysis using an Operational Perspective

As discussed in Section 2.1, it is not uncommon in OT to choose a grand mean of the response as the parameter of interest. Within an ANOVA framework (whether used implicitly or explicitly), however, this choice can be problematic. Averaging across factor levels assumes both an ordering of levels and a linear relationship between those levels and the response, neither of which is implicit in the ANOVA framework. This would have the effect of evaluating a measure in an operational environment that may not exist. Averaging across factor levels also has the effect of artificially reducing operational variability (see Figure 2.1): if a system performs well above ϕ_0 in one operational environment and just below ϕ_0 in all others, it is possible to evaluate the measure as met when using a grand mean, even if ϕ_0 would not be met in most operational environments. Similarly, if a system performs just above ϕ_0 in most environments and well below ϕ_0 in one, it is possible to evaluate the measure as not met when using a grand mean.

In contrast, "mission sets" (a combination of factor levels) that represent operational environments are proposed, providing a method to obtain a summary statistic based on performance across operational environments without artificially reducing the operational variance. This is accomplished by considering the joint distribution of mission sets and mission means, and then marginalizing over mission sets to obtain a mixture distribution of mission means. This grounds measure evaluations in actual results from representative operational environments, and prevents the previously addressed problems that can arise when the variability is artificially reduced. We refer to this method as the Bayesian mission mean approach. At the end of the next sub-section, once all components of the proposed method in this chapter have been introduced, equations (2.1)–(2.3) demonstrate the difference between a frequentist grand mean approach, a Bayesian grand mean approach, and a Bayesian mission mean approach.

Mission sets are drawn based on the anticipated likelihood of encountering levels established during factor prioritization. These mission sets do not represent the space of all mission sets, but rather those used to evaluate the system. In an ANOVA model, to construct a single mission set, one level from each factor is drawn from a corresponding categorical distribution; the number of categories (k) in the categorical distribution is defined by the number of levels for the given factor, and the probabilities $(p_1, \ldots, p_k)$ are defined by anticipated likelihood of encountering each level. For example, consider a factor with two levels (e.g. α_i, $i = 1, 2$); the index defining which level is included in a mission set would be drawn from a Cat(2, p_1, p_2). An example relating mission sets to an operational environment is presented in Section 2.3.

Incorporating mission sets into the analysis provides a novel way of conducting OT when the parameter of interest is a summary statistics that will be used to evaluate a measure. By approaching analysis over the marginalized mission sets, the operational environment is defined more rigorously and provides more actionable in-

formation. When using a grand mean, the small variance of ϕ masks the complexity of the operational environment. Using the marginalized mission space provides more information about the variability in the parameter of interest within the operational environments the system will be used in. By accounting for the range of possible outcomes, the operational environment is being explicitly used to evaluate system performance—and therefore explicitly used to make decisions about system procurement. This more rigorous definition of the operational environment more closely aligns with the goal of OT: determining if the user can accomplish the mission when employing the system in an operational environment.

2.2.3 Operational Testing using Bayesian Analysis

Finally, the shift to a Bayesian perspective allows testers to calculate the direct probability of the parameter of interest obtaining the threshold value, $\Pr_{\phi|X}(\phi \geq \phi_0)$, after a test has been completed and some data, X, have been obtained. To obtain this probability, first a joint posterior distribution is numerically approximated using traditional MCMC sampling methods in software such as R or OpenBUGS and then mission sets are drawn. Once the mission sets and the posterior distribution on the model parameters are obtained, they can be used to induce a distribution on ϕ. This is accomplished by using the mission sets to mix the mission means from the posterior distribution. This induced distribution can then be used to calculate the direct probability that the parameter of interest is greater than the threshold value across all mission sets. Given this, a certainty threshold value (θ_T) is selected; θ_T is a probability which is used to express how much certainty is required before stating that $\phi > \phi_0$. Instead of the current practice of comparing ϕ to ϕ_0, this method uses θ_T to re-define how how a measure is evaluated:

– If $\Pr_{\phi|X}(\phi \geq \phi_0) > \theta_T$, the measure would be evaluated as met.

– If $\Pr_{\phi|X}(\phi \geq \phi_0) \leq \theta_T$, the measure would be evaluated as not met.

Using this construct, equations (2.1) - (2.3) demonstrate the difference between a grand mean (GM) approach and a mission mean (MM) approach. Let $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (the mean of the data) and let σ be the standard deviation of the data. Let μ_j be a normally distributed random variable that represents the mean for the j^{th} mission set, $j = (1, \ldots, m)$, and let p_j be the probability of seeing the j^{th} mission set ($\sum_{j=1}^{m} p_j = 1$). Finally, let δ be a vector from a Multinomial$(1, p_1, \cdots, p_{m-1})$, where δ_j is the j^{th} element of δ. Equation (2.1) is a frequentist evaluation of the grand mean, using a traditional statistical test (instead of simply $\bar{y} > \phi_0$); equation (2.2) is a Bayesian evaluation of the grand mean; and equation (2.3) is a Bayesian evaluation of the mission mean:

$$\frac{\bar{y} - \phi_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\theta_T} \tag{2.1}$$

$$\Pr(\phi_{GM} > \phi_0) \geq \theta_T \tag{2.2}$$

$$\Pr(\phi_{MM} > \phi_0) \geq \theta_T \tag{2.3}$$

where $\phi_{GM} = \sum_{j=1}^{m} \mu_j p_j$ and $\phi_{MM} = \prod_{j=1}^{m} \mu_j^{\delta_j}$. Therefore, ϕ_{GM} is a weighted average of random variables and ϕ_{MM} is a random selection of random variables.

2.3 Electric Semi-Truck Example from a Bayesian Perspective

In this section, the electric semi-truck example will be explored from a Bayesian perspective, evaluating the electric semi-truck using the Bayesian mission mean method. Overlaying the Bayesian framework on the current process, an ANOVA

model with main effects and two-way interactions (excluding wind) is selected and explicitly used. Table 2.2 updates Table 2.1, accordingly, and defines the ANOVA model parameters. The ANOVA model is then:

$$y_{ijklmp} = \mu_{ijklm} + \epsilon_{ijklmp} \qquad \begin{cases} i = 1, 2 \\ j = 1, 2 \\ k = 1, 2, 3 \\ l = 1, 2 \\ m = 1, 2 \\ p = 1, \ldots, 5 \end{cases}$$

where

$$\mu_{ijklm} = \eta + \alpha_i + \beta_j + \omega_k + \gamma_l + \delta_m + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\alpha\delta)_{im} + (\beta\gamma)_{jl} + (\beta\delta)_{jm} + (\gamma\delta)_{lm}$$

and p represents the replicate. Furthermore, using a reference cell model, η is the baseline parameter representing the first level of each factor. The constraints of this reference cell model are all model parameters at the first level, or with at least one factor at the first level for interactions, are equal to 0. Assuming that the errors are *iid* and normally distributed with a mean of 0 and a variance of $\frac{1}{\tau}$ (where τ is referred to as the precision), the distribution of miles traveled within each group can be written as

$$y_{ijklmp} | \mu_{ijklm}, \tau \overset{iid}{\sim} N\left(\mu_{ijklm}, \frac{1}{\tau}\right).$$

Next, independent Normal priors that took into account the plausible range of parameter values were selected. For example, without data relating to the electric semi-truck, testers believed changing the baseline parameter (η) from hilly terrain ($i = 1$) to flat terrain ($i = 2$) would increase the average number of miles traveled by 50 miles. They further determined that it was unrealistic to expect an increase of more than 250 miles, given this change. For this example, placing a variance of 100^2

on the distribution of the flat terrain parameter, α_2, would result in an approximate 95% Normal interval of change-in-average-miles-traveled from -150 (two standard deviations below the mean of 50) to 250 (two standard deviations above the mean). Testers believed that a change outside of this span was extremely unlikely; therefore, they selected an independent Normal(50, 100^2) prior for α_2. Similar assessments were made for each unconstrained model parameter and the selected priors can be found in Table 2.2. None of the means for these priors are 0, which reflects the belief that these factor levels have an effect on the response variable—the reason these factor levels were selected. Furthermore, 0 is contained in either two standard deviations below or above the mean, allowing for the possibility that the model parameters may not be significant, or may have a different direction then believed. Model parameters that are constrained to 0 result in priors that are degenerate at 0, which can be seen in Table 2.2.

Factor	Levels	Effect	LoE*	MP*	Prior
Terrain	Hilly Flat	High	50% 50%	α_i	$\Pr(\alpha_1 = 0) = 1$ $p(\alpha_2) \sim \mathrm{N}(50, 100^2)$
Temperature	Hot Moderate	Medium	4/9 5/9	β_j	$\Pr(\beta_1 = 0) = 1$ $p(\beta_2) \sim \mathrm{N}(50, 50^2)$
Wind	Good Moderate Poor	Medium	1/3 1/3 1/3	ω_k	$\Pr(\omega_1 = 0) = 1$ $p(\omega_2) \sim \mathrm{N}(-25, 50^2)$ $p(\omega_3) \sim \mathrm{N}(-50, 50^2)$
Payload Type	Refrigerated Non-Refrigerated	High	50% 50%	γ_l	$\Pr(\gamma_1 = 0) = 1$ $p(\gamma_2) \sim \mathrm{N}(100, 100^2)$
Weight	Heavy Light	High	50% 50%	δ_m	$\Pr(\delta_1 = 0) = 1$ $p(\delta_2) \sim \mathrm{N}(100, 100^2)$

* Likelihood of Encountering (LoE) and Model Parameter (MP)

Table 2.2: Factor Prioritization for Electric Engine

Given that the electric semi-truck had proceeded to an operational test, testers believed that a mean of 400 miles (the threshold) would be reasonable for the baseline

parameter (η, which represents hilly terrain, hot temperature, good wind, refrigerated payload type, and heavy weight); they also found it extremely unlikely that the engine could travel 600 miles on one charge in the baseline case. Therefore, a Normal(400, 100^2) prior was selected for η. This places 95% of the probability between 200 and 600 miles. Additionally, a Normal(0, 100^2) prior was selected for all two-way interactions to allow for two-way interactions to increase the response, decrease it, or have no effect on it. Finally, a Gamma$(0.0001, 0.0001)$ reference prior was selected for the precision, τ.

As changes to the current process did not influence the experimental design selection, the same 2^4 full factorial with five replicates from Section 2.1.2 was selected. Additionally, a certainty threshold of $\theta_T = 0.8$ was selected.

At the end of test execution, a posterior distribution is obtained. For each MCMC draw from the posterior distribution of model parameters, a mission set is also drawn, as discussed in Section 2.2.2. For instance, to obtain the terrain index, i, for a mission set: i is drawn from a Cat(2, 0.5, 0.5) distribution, where obtaining a 1 would correspond to a hilly terrain (α_1) and a 2 would correspond to a flat terrain (α_2). To demonstrate how this relates to an operational environment, consider if the only two factors were terrain (hilly and flat) and temperature (moderate and hot). If the drawn mission set is flat / hot, this representation could correspond to an operational environment in Florida, for example.

After inducing a distribution on the mean miles traveled and marginalizing over the mission sets, $\Pr_{\phi|X}(\phi \geq 400)$ can be calculated and compared to $\theta_T = 0.8$ to evaluate the measure. The posterior distributions using the 21 data sets were calculated using a Gibbs sampler in R, with 80,000 posterior samples (after examining the trace plots, and removing the burn-in samples). Results from the analysis for each data set can be found in Table 2.3, along with the results from the current OT process described in Section 2.1.2.

		Proposed Method: Bayesian Mission Mean		Current Method: Grand Mean	
Data Set	η	$\Pr_{\phi\|X}(\phi \geq 400)$	Measure is...	Mean	Measure is...
1	343	0.7856	Not Met	481.46	Met
2	345	0.7968	Not Met	483.46	Met
3	347	0.8059	Met	485.46	Met
4	349	0.8134	Met	487.46	Met
5	351	0.8212	Met	489.46	Met
6	353	0.8302	Met	491.46	Met
7	355	0.8380	Met	493.46	Met

(a) Error Transformation 1 (Smallest Variance)

		Proposed Method: Bayesian Mission Mean		Current Method: Grand Mean	
Data Set	η	$\Pr_{\phi\|X}(\phi \geq 400)$	Measure is...	Mean	Measure is...
8	343	0.7812	Not Met	480.51	Met
9	345	0.7879	Not Met	482.51	Met
10	347	0.7959	Not Met	484.51	Met
11	349	0.8018	Met	486.51	Met
12	351	0.8096	Met	488.51	Met
13	353	0.8172	Met	490.51	Met
14	355	0.8248	Met	492.51	Met

(b) Error Transformation 2

		Proposed Method: Bayesian Mission Mean		Current Method: Grand Mean	
Data Set	η	$\Pr_{\phi\|X}(\phi \geq 400)$	Measure is...	Mean	Measure is...
15	343	0.7802	Not Met	479.42	Met
16	345	0.7885	Not Met	481.42	Met
17	347	0.7916	Not Met	483.42	Met
18	349	0.7998	Not Met	485.42	Met
19	351	0.8062	Met	487.42	Met
20	353	0.8123	Met	489.42	Met
21	355	0.8195	Met	491.42	Met

(c) Error Transformation 3 (Largest Variance)

* Error Transformation definitions are in Appendix B.

Table 2.3: Bayesian Mission Mean Approach for Various ηs and Error Transformations, Compared to a Grand Mean Approach

Using the current grand mean method, all 21 data sets resulted in the measure being evaluated as met; in contrast, only 12 data sets result in the measure being evaluated as met when using the Bayesian mission mean method. This is due to the current grand mean method artificially reducing the variability of actual performance in operational environments. In contrast to the grand mean, by marginalizing over the mission space, the Bayesian mission mean method takes into account the operational variability and restricts the evaluation focus to system performance in the operationally representative environments where testing is conducted. To illustrate this, consider Figure 2.1. For ease of comparison, we are showing the Bayesian grand mean and Bayesian mission mean because they are both functionals of the same unknown model parameters. The frequentist grand mean is a functional of the data, and less directly comparable.

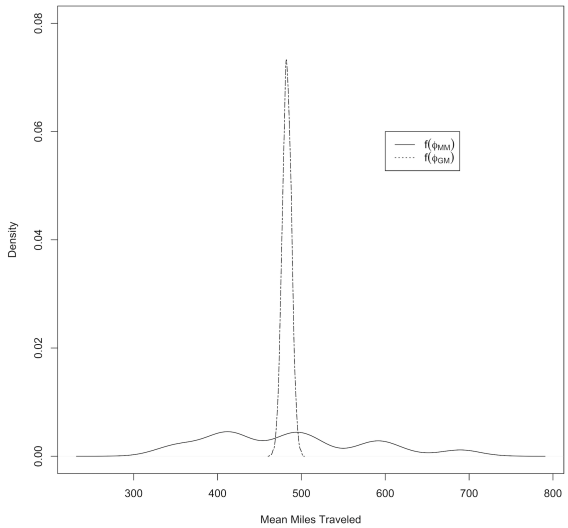

Figure 2.1: Data Set 2 Densities for ϕ_{GM} and ϕ_{MM}

While model parameter estimates are not the focus of our method for evaluating measures, should knowing their values be a secondary objective, they can still be obtained. An example of model parameter estimates is provided in Table 2.4. This table also gives the reader a sense of how the priors, posteriors, and true values relate.

It can be seen that, although the posteriors remain somewhat diffuse, the true values are all within a central 95% probability interval on the posterior.

Model Parameter	Value for Data Generation	Prior E(·)*	Posterior E(·)*	Prior sd(·)*	Posterior sd(·)*
η	347	400	352.75	100	18.82
α_2	50	50	60.07	100	21.30
β_2	15	50	15.49	100	19.73
ω_2	0	-25	-6.62	50	15.27
ω_3	-5	-50	-4.01	50	12.93
γ_2	75	100	66.27	100	21.08
δ_2	50	100	56.92	100	21.66
$(\alpha\beta)_{(22)}$	10	0	-12.61	100	21.16
$(\alpha\gamma)_{(22)}$	50	0	38.54	100	21.69
$(\alpha\delta)_{(22)}$	25	0	17.15	100	22.22
$(\beta\gamma)_{(22)}$	50	0	65.56	100	21.23
$(\beta\delta)_{(22)}$	25	0	15.80	100	21.32
$(\gamma\delta)_{(22)}$	25	0	22.22	100	23.22
τ	See Appendix B	1	0.0004142	100	0.0000707

* E(·) is the expectation of the model parameter and sd(·) is the standard deviation.

Table 2.4: Model Parameter Estimates with $n = 80$ Observations for Data Set 3

An advantage of this method, in addition to evaluating the measure over the marginalized mission sets, is that further information can be provided regarding specific mission sets. Therefore, not only does this method address the goal of OT, but it can also be used to inform planning by users in the operational environment. For instance, if the transport company's logistical division wanted more information about the operational environment represented by the baseline parameter for planning purposes, this method can address that question. Instead of randomly generating mission sets over the space of mission sets, the baseline mission would be used for every posterior draw already obtained.

Mission Set*	Posterior Expectation of μ_{ijklm}	Posterior Standard Deviation	$\Pr_{\phi\|X}(\phi > 400)$
{11111}	352.75	18.82	0.0063
{22111}	415.71	19.10	0.7955
{12112}	440.97	19.76	0.9808
{12311}	364.24	19.66	0.0355
{22222}	691.55	19.81	$>0.9999^{\dagger}$

* A mission set is defined by the indices of μ in the ANOVA model, $\{ijklm\}$.
† Computationally, this is numerically indistinguishable from 1, but not actually 1.

Table 2.5: Subset of Mission Set Estimates for Data Set 3

Implementing this method for planning purposes, Table 2.5 provides a subset of mission sets and the posterior expectation of the number of miles traveled from the corresponding induced distribution; this table also highlights that not every posterior expectation need be above 400 miles for the measure to be evaluated as met.

Chapter 3

Adaptive Operational Testing

By shifting to a Bayesian Framework in Chapter 2, testers were able to obtain direct probabilities, instead of indirect probabilities. In addition to providing direct probabilities, the Bayesian framework also allows for *predictive probabilities*, which can be obtained from a predictive distribution of future observations, given the data already obtained. A predictive distribution is a distribution of unobserved future observations, given the observations already seen; after obtaining the predictive distribution, estimates of interest can be calculated, to include predictive probabilities. For example, the probability that a future observation will exceed a given value when it has yet to be seen can be calculated within a Bayesian framework. Frequentists can provide a point estimate for a future observation, or a prediction interval for that point estimate, but "[p]robabilities of future observations are not possible in a formal frequentist approach" (D. A. Berry 1993).

Furthermore, the Bayesian paradigm provides a flexible approach to analysis, which frequentist methods require more effort to achieve. Within the Bayesian framework, the constant updating of information is easily accomplished. As data accumulates, inferences can be updated and new conclusions can be made; therefore,

each new data point is an update to the current belief. As a result, interim results can then become final results at any time during a test. Within the frequentist paradigm, conducting interim analysis with intent to stop test early requires adjustments to p-values; this is not the case for conducting Bayesian interim analysis. In a Bayesian analysis, inferences are not affected by interim analyses or the reason a test was stopped as a result of the likelihood principle and the lack of dependence on an experimental design (D. A. Berry 1993; Zang and Lee 2014; D. A. Berry 1987). This flexibility in Bayesian analysis provides a natural framework for allowing operational testers to answer the question "how much testing is enough?" during test execution. If the question is satisfactorily answered earlier in testing, stopping the test would provide both cost and schedule savings. The idea of constant updating in the Bayesian framework can be expressed in the following manner:

> ...Bayesian inference is not merely data analysis. Bayes' theorem is a formalism for learning: that's what I thought before, this is what I just saw, so here's what I now think—and I may change my views tomorrow.
> - Donald A. Berry, A Case of Bayesianism in Clinical Trials (1993)

This idea is not only a central tenet of Bayesian statistics, but it is also the underlying concept of a military theory for decision making: John Boyd's "OODA Loop".

As described by Hammond, Boyd considered the OODA Loop (short for the Observe, Orient, Decide, Act Loop) to be "a composite of how we think and learn, the source of who we are, and the potential we possess... a shorthand for life itself, a model for how we think" (Boyd 2018). The OODA Loop—the iterative process of folding new information into a constantly updated view of the world from which actions and feedback naturally flow—is not limited to a military training, but is now taught across a variety of professions (Boyd 2018). The loop begins with "observation": assessing the current environment. Next, "orientation" considers previous

experience (traditions, education, personal experiences, etc.), and synthesizes it with the information obtained from observation. This allows for various courses of action to be established. A "decision" is then made regarding which course to select, based on some set of decision criteria (e.g., which course of action has the least risk). Finally, "act" implements the course of action decided upon. In its typical representation (which is considered an over-simplification, c.f. Hammond's Appendix on the OODA Loop, Boyd 2018), the OODA Loop is presented in a circular pattern—each piece of the loop feeding the next, until act ultimately feeds into observation and the loop begins again (Boyd 2018). While the OODA Loop was originally conceptualized for tactical and strategic engagements, it can also be seen as an analytical tool (Boyd 2018). When used in this manner, Bayesian methods can be seen as the statistical manifestation of the OODA Loop. An experiment is observed and data are obtained (such as in operational testing (OT)); next, a prior based on an individual's beliefs (such as the belief that developmental testing (DT) data is commensurate with OT data) is then synthesized with the experimental data (orientation), which allows for a decision to be made about what that synthesis means (e.g., is the system mission capable?; is more testing required?). Having decided, an action is then taken (e.g., procure the system; accomplish more testing). Therefore, not only does the Bayesian perspective provide statistical capabilities that are absent in the current OT paradigm, it also aligns with an important military theory for decision making.

Having transitioned to a Bayesian framework, Section 3.1 begins with a literature review of specific work within clinical trials using predictive probability. Section 3.2 broadens this clinical trial work to OT and extends that work to a fully Bayesian method for a continuous response with an ANOVA structure. This section develops a method for adaptive OT by considering the case where interim data are obtained and consideration can be given to stopping a test early. Using the electric semi-truck example, Section 3.3 implements the proposed method.

3.1 Clinical Trials

Bayesian methods have become increasingly popular in clinical trials due to the flexibility and natural interpretation of the analysis such methods provide (D. A. Berry 1993; Zang and Lee 2014; M. Liu and Dressler 2018). This section focuses on the second of three phases in clinical trials, phase II, where the goal is to understand if the drug is effective—typically by comparing a binary response to a standard (be it the current standard of effectiveness, or historical standards) (Zang and Lee 2014; Yin, N. Chen, and Lee 2012). During a trial, monitoring incoming data with the intent of deciding whether sufficient information has been obtained to make such comparisons before the end of a trial is referred to as *interim analysis* and is considered an adaptive testing method. One method for employing Bayesian interim analysis within clinical trials is to incorporate predictive probabilities of eventual trial success.

Lee and D. D. Liu (2008) discuss how interim analysis can use predictive probabilities to end trials early when the response is binomial. With increasing interest in Bayesian adaptive methods in trials, the on-going multi-drug trial I-SPY 2 (Barker et al. 2009) and a completed drug trial adding trastuzumab to chemotherapy (D. A. Berry 2005) are two examples of clinical trials with binomial responses that have successfully incorporated predictive probabilities. Furthermore, the latter trial validated the utility of Bayesian predictive probability within a frequentist design framework (D. A. Berry 2006).

Geisser and W. Johnson (1994) detail how predictive probabilities can be used in interim analysis when the response is continuous, but discuss the computational difficulty and ultimately offer distributional approximations instead. Dmitrienko and Wang (2006) explore a continuous response for a clinical trial that compares two treatments. However, they acknowledge that their work is not a fully Bayesian ap-

proach, and state that "a fully Bayesian solution in the case of normally distributed [responses] is extremely complex from a computational perspective and will be difficult to use in practice." Recent work by M. Liu and Dressler (2018) extend Lee and D. D. Liu's work to a simple case with a single treatment where a closed-form solution can be obtained for a continuous response. Furthermore, they state that more research is required for posterior distributions that do not have a closed-form. Zhou et al. (2018) outline a general framework for predictive probability when a closed-form solution is not available, but acknowledge the computational complexity and do not implement the framework. Our understanding of what Zhou et al. and M. Liu and Dressler are describing is that they are using the term "closed-form" to describe a recognizable distribution.

The method presented in this chapter leverages this work with predictive probabilities in clinical trials to stop drug trials early—specifically, Lee and D. D. Liu's work with predictive probabilities. The traditional Bayesian goal of a phase II clinical drug trial is to determine whether the probability of the response rate, p, being greater than some hypothesized value, p_0, is above some pre-defined probability threshold, θ_T. When interim data can be obtained, a new goal can be established, answering the question: how likely is it that a conclusive statement will be made at the end of the trial, given the data that have already been seen.

Within the Bayesian framework, predictive probabilities for future observations are a natural way to decide if a test should be stopped early. In this framework, "we can condition on future results, evaluate their consequences, and average with respect to these probabilities" (D. A. Berry 1993). Therefore, decisions to stop test can be based on the predictive probability of a successful completion of the trial (PP). Lee and D. D. Liu (2008) suggest the use of probabilities θ_L and θ_U to decide whether a trial should be stopped or not. If PP is less than θ_L, the trial can be stopped early because testers are confident they would make an assessment of drug futility at the

end of the trial. If PP is greater than θ_U, similarly the trial can be stopped early because they are confident they would make an assessment of drug efficacy at the end of the trial. If PP is between θ_L and θ_U, the trial continues because there is not enough data to make a conclusion regarding the drug (Lee and D. D. Liu 2008).

At some intermediate point during a drug trial, let x be the number of successful responses that have been obtained and Y be the number of successful responses that have yet to be obtained. Then PP can be written as:

$$PP = \Pr_{Y|x}(Y : \Pr(p > p_0|x, Y) > \theta_T) \tag{3.1}$$

$$= E\{I[\Pr(p > p_0|x, Y) > \theta_T]|x\} \tag{3.2}$$

(Lee and D. D. Liu 2008). Considering the right hand side of equation (3.1), $p > p_0$ is the inequality that determines if the response rate, p, is greater than the response rate the drug must obtain to be considered effective, p_0. $\Pr(p > p_0|x, Y)$ expresses the probability of this inequality being true, given the data that have and have not seen, which is then used to conduct a statistical test against a probability threshold, θ_T. Finally, the probability of a trial being a success is calculated. This statistical test is reminiscent of the one used in Chapter 2 to evaluate the measure; however, not all observations have been observed for this statistical test. Therefore, the probability of a test being a success (the expectation of Y over the predictive distribution of $Y|x$) is calculated.

To illustrate PP, a graphical representation has been developed and can be found in Figure 3.1, derived from an example by Lee and D. D. Liu (2008). In the example, the goal is to determine if $\Pr(p > 0.6|x, Y) > \theta_T$. No more than 40 patients will be seen, of which 23 have already been seen and 16 have had a successful response; therefore, $x = 16$ and $Y \in (0, 1, \ldots, 17)$. The prior on p is $\text{Beta}(0.6, 0.4)$ and $Y|p$ is $\text{Bin}(17, p)$. When Y is univariate (or a univariate summary can be obtained), its joint posterior with the response rate can be graphically represented. The diagonal

solid line is the $(1-\theta_T)^{th}$ quantile of the posterior distribution of $p|Y$, conditional on each possible Y. The horizontal dashed line represents when $p = p_0$. The intersect of the solid and dotted line provides a cutoff value for Y; Y's greater than this cutoff value represent the Y values for which the $(1-\theta_T)^{th}$ quantile of the distribution is greater than p_0. Finally, the shaded portion indicates those values of Y that would result in a favorable conclusion about the test if they were seen. The probability under this shaded portion—the probability for Y under its predictive distribution, see equation (3.1)—corresponds to the probability that a favorable test result would be found if the test was run through completion. Note that in Figure 3.1 Y can only take on integer values; the density curves displayed are an artifact of the plotting method in R, and were maintained to help visualize the relative density around each integer value of Y.

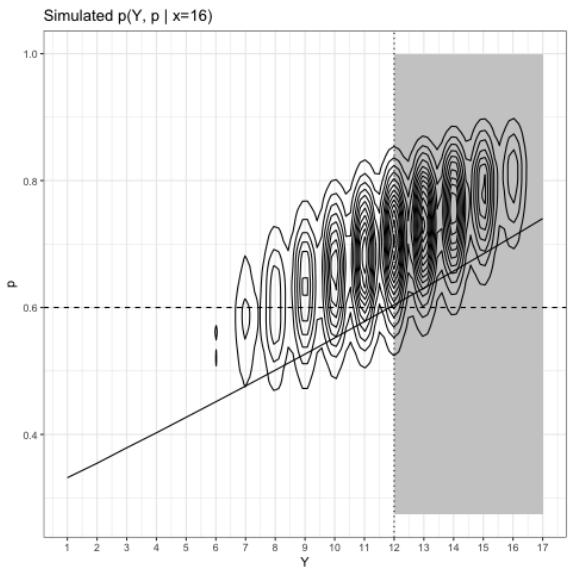

Figure 3.1: Graphical Representation of PP, $\theta_T = 0.1$

3.2 Adaptive Operational Testing

There are many differences between clinical trials and OT that must be addressed in order to implement PP in OT. One difference is that the data from OT can have a diverse set of distributions for responses and priors that are appropriate, unlike the binomial distribution with a beta prior that typically works well for the commonly used binary response in phase II clinical trials (Zang and Lee 2014; Yin, N. Chen, and Lee 2012; Thall et al. 2007). While this makes OT flexible, it also makes it more complex. Additionally, a fully Bayesian approach to a continuous response with an ANOVA structure leads to a more complex sampling process, where covering the space of possible outcomes becomes more challenging. Furthermore, phase II clinical drug trials are inherently comparative in nature and are interested in how effective a drug is for specific patient and disease profiles. This is in contrast to this research, which uses an experimental design to then characterize the performance of a system across factors and levels. The proposed method addresses the aforementioned differences and extends the clinical trial work with predictive probability into the OT framework.

In developing a method for adaptive OT, the same test planning process as detailed in Section 2.2.2–2.2.4 is used. Next, PP is incorporated. The analogous equations to (3.1) and (3.2) from Section 3.1 for OT with a continuous response are:

$$PP = \Pr_{Y|X}(Y : \Pr(\phi > \phi_0|X, Y) > \theta_T) \tag{3.3}$$

$$= E\{I[\Pr(\phi > \phi_0|X, Y) > \theta_T]|X\}, \tag{3.4}$$

where X is the set of responses that have been obtained and Y is the set of responses that have yet to be obtained.

Tolerance parameters for stopping test early, θ_L and θ_U, are also incorporated, providing an avenue for testers to determine if they have enough data to make conclusions about a measure without completing every test event (Lee and D. D. Liu

2008). If PP is less than θ_L, the test can be stopped early because testers are confident they would evaluate the measure as not met at the end of the test. If PP is greater than θ_U, similarly the test can be stopped early because they are confident they would evaluate the measure as met at the end of the test. If PP is between θ_L and θ_U, the test continues because there is not enough data to make a conclusion regarding the measure (Lee and D. D. Liu 2008).

Testers should choose levels for θ_L and θ_U based on their tolerance for incorrectly accepting or rejecting a system, and those values may vary between systems depending on the consequences for stopping a test incorrectly. For instance, consider body armor; due to the life-saving nature of the system, the consequences for accepting the system incorrectly may outweigh the consequences for rejecting the system incorrectly. Tests in such an instance may only be allowed to stop when testers become confident that the system will not obtain the threshold value, but will not be stopped early otherwise. Therefore, to evaluate a measure as met, it would require the full experimental design to be accomplished; however, not all runs would need to be accomplished to evaluate the measure as not met. In this instance, θ_U would be set to 1. While rare that θ_L would be set to 0 or θ_U would be set to 1 in OT, it is important to weigh the pros and cons of the potential decision before test execution.

After establishing θ_L and θ_U, the number of observations required to be seen before calculating PP (referred to as n_f) should be established. While PP can be calculated starting at $n_f = 0$, if it is calculated when only a few observations have been obtained, it is possible to make the wrong decision due to a lack of information (Lee and D. D. Liu 2008). Furthermore, it is known that using reference priors can result in high stopping rates that may be undesirable for testing (Saville et al. 2014). Due to these issues, we do not recommend using $n_f = 0$ when using references priors. Instead, we recommend choosing n_f by considering both the coverage of level combinations and the number of observations informing the posterior for each model

parameter before calculating PP, which can be done by looking at the design matrix. For example, n_f could be the number of observations after at least four observations for each parameter in the model have been observed with at least 80% of the linear combinations having been seen. Therefore, n_f will be a function of the order in which test events are run based on the randomized design matrix; two different randomized design matrix under the same experimental design will often give two different n_fs. As the criterion for decisions about a measure is based on a summary of performance across the linear combinations, not all linear combinations have to be seen before calculating PP because of the linear structure of the model. PP can be calculated as long as enough information has been obtained for the model parameters and the linear combinations are sufficiently covered to reduce collinearity between parameter estimates (as frequently occurs with D-optimal designs). This method provides an efficient use of test resources when decisions are based on a summary across the group means; this in turn allows testers to estimate the group means well enough to make a decision without expending additional resources to focus on estimating the group means with high precision.

After establishing θ_L, θ_U, and n_f, testing begins and PP can be calculated after n_f observations are obtained. Of note, the frequency of interim analysis does not need to be done at regular intervals. Furthermore, while PP can be calculated after every new observation, given computational requirements of the method, it may only be practical to do interim analysis and calculate PP after sets of observations have been obtained. Finally, the frequency of obtaining interim data, as well as the precision of that data, will be defined by the system and the data collection process. After comparing PP to θ_L and θ_U, a decision to stop testing early or not can be made. If PP never allows the test to stop early, and all observations are seen, the analysis returns to using posterior probability (Chapter 2) and $\Pr_{\phi|X}(\phi \geq \phi_0) > \theta_T$ is used to evaluate the measure.

A flow chart of this method can be found in Figure 3.2. In this flow chart, let n be the total number of test events from an experimental design and let n_{obs} be the number of observations that have been seen ($n_{\text{obs}} \leq n$). Shaded boxes are the end points of the method.

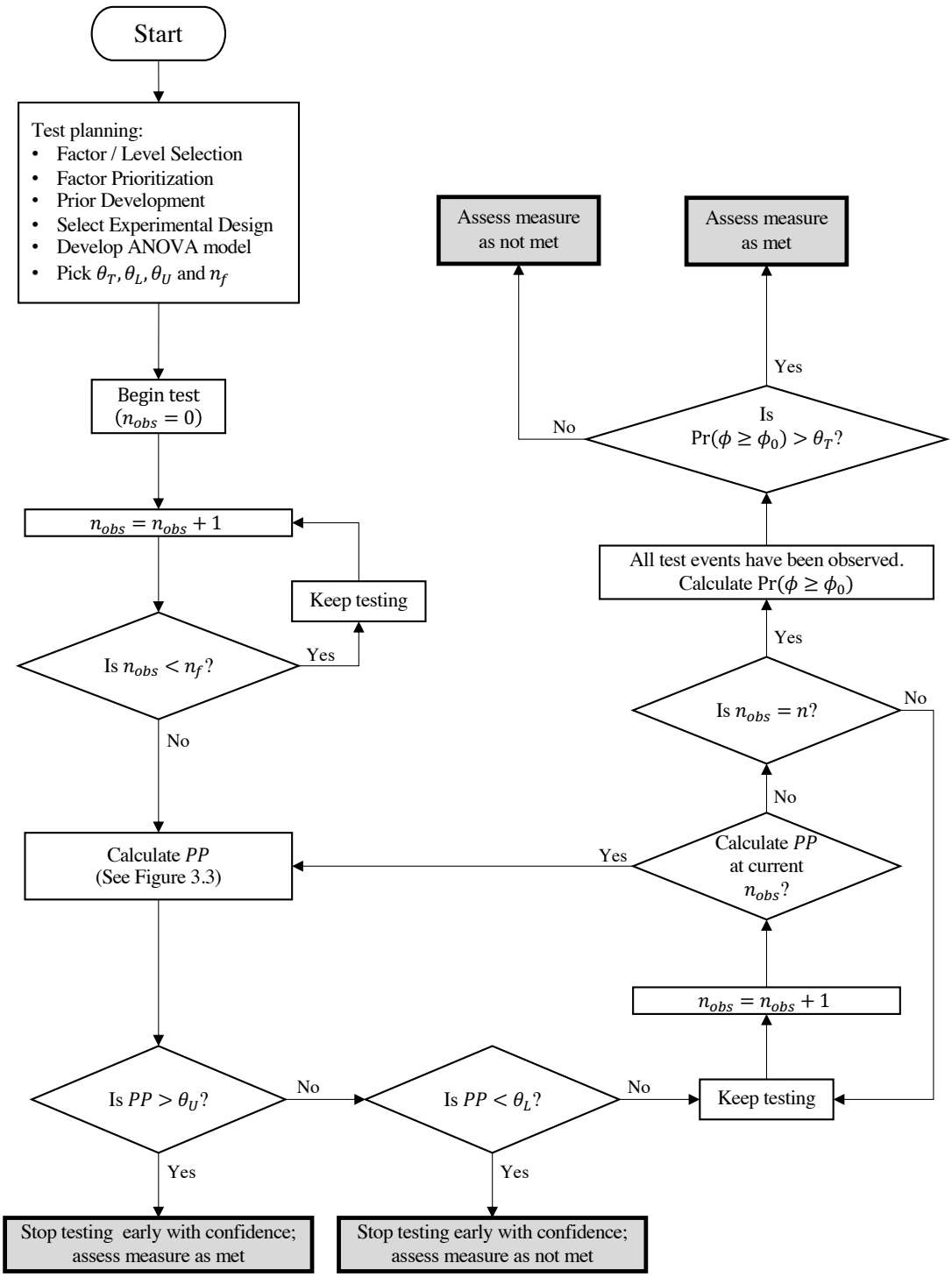

Figure 3.2: Adaptive Operational Testing Process

Using posterior probability in the case where all the data have been obtained (Chapter 2) can be done using traditional sampling methods; the sampling method required to obtain PP is more complex. In contrast to a Gibbs sampler (or Metropolis-within-Gibbs sampler), a two-stage sampling method is required. The sampling method proposed in this section aligns with Zhou et al.'s general framework (Zhou et al. 2018), and provides a concrete process for practitioners to implement. Traditionally, the interest is in a sample from the joint posterior distribution of Y and the model parameters; however, incorporating θ_T in PP necessitates sampling from a conditional distribution of model parameters, given Y, for each possible Y yet to be seen. Therefore, the nested sampler determines the conditional posterior distribution of the model parameters, given Y, and the outside sampler determines the marginal posterior distribution of Y (also commonly called the predictive distribution for Y). While this is a non-traditional sampling method, this approach still provides a sample from the joint posterior distribution of Y and the model parameters, as the joint posterior distribution is proportional to product of the marginal posterior distribution of Y and the conditional posterior distribution of the model parameters given Y. This construct requires careful consideration when choosing how many Ys should be sampled and how many conditional posterior samples for that Y need to be obtained to avoid overly burdensome computational requirements while still covering the space of model parameters and possible Ys adequately.

The algorithm for this process can be seen in Figure 3.3. In this algorithm, let ξ represent the set of K model parameters and ξ_k represent the kth parameter $(k = 1, \ldots, K)$. Let Y be the set of unseen observations and let ϕ be the parameter of interest, which is a function of the response. Let i be the ith iteration from the outer sampler and b_i be the number of burn-in samples for the outer sampler; then $i = 1, \ldots, b_i + n_i$. Finally, let j be the jth iteration from the nested sampler and b_j be the number of burn-in samples for the nested sampler; then $j = 1, \ldots, b_j + n_j$.

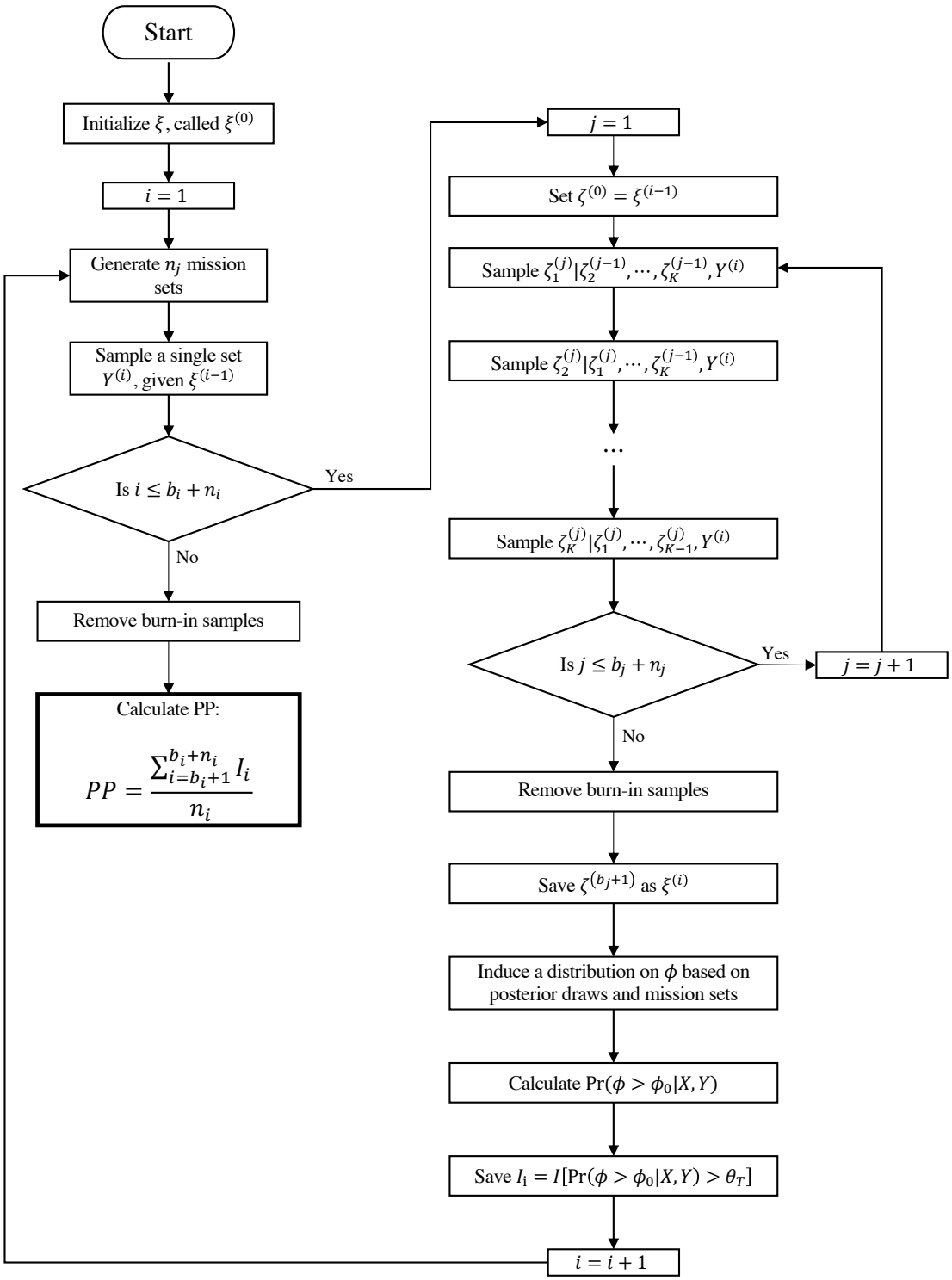

Figure 3.3: Two-Stage Sampling Algorithm

3.3 Electric Semi-Truck Example

Returning to the electric semi-truck example, the same set-up and experimental design is used as in Section 2.3.2. Additionally, tolerance values of $\theta_L = 0.05$ and $\theta_U = 0.95$ were selected after testers considered the implications of the choice. Finally, PP would be calculated starting at $n_{\text{obs}} = 45$ (where $45 > n_f$). Then, PP for this example is:

$$\begin{aligned} PP &= \Pr_{Y|X}(Y : \Pr(\phi \geq 400|X, Y) > 0.80) \\ &= E\{I[\Pr(\phi \geq 400|X, Y) > 0.80]|X\}. \end{aligned}$$

Once $n_{\text{obs}} = 45$, testers calculated PP. Selecting the means of the prior distributions to initialize each model parameter, let $\mu^{(0)}_{ijklm}$ represent the ANOVA model with the initialized model parameters and $\tau^{(0)}$ represent the initialized precision. Beginning with the outer sampler, the remaining 35 future observations (the set of remaining $y^{(1)}_{ijklmp}$s) were then sampled from a $N\big(\mu^{(0)}_{ijklm}, \frac{1}{\tau^{(0)}}\big)$ based on the remaining rows within the design matrix. These observed and sampled responses make up all 80 runs for the experimental design, which are then used in the nested sampler to obtain the conditional posterior distribution of the model parameters, given the observations that have and have not been seen.

For each draw from the conditional posterior distribution of model parameters, a mission set is also drawn, as detailed in Chapter 2. Using the mission sets and the conditional posterior distribution of model parameters, a distribution can be induced on the mean number of miles traveled on one charge. After marginalizing over the mission sets, the $\Pr_{\phi|X}(\phi \geq 400) > 0.80$ can be calculated and stored for the set of $y^{(1)}_{ijklmp}$s sampled. This conditional posterior is also used to update the outer sampler and generate a new set of 35 future observations (the set of remaining $y^{(2)}_{ijklmp}$s) from a $N\big(\mu^{(1)}_{ijklm}, \frac{1}{\tau^{(1)}}\big)$. After obtaining a sufficient sample from both distributions, PP is calculated and compared to θ_L and θ_U to determine if testing could stop early.

PP was calculated in R using the two-stage sampling algorithm detailed in Figure 3.3, based on $n_j = 18,000$ nested samples and $n_i = 1,000$ outer samples (after examining both ACF and trace plots, and removing burn-in samples). To illustrate how this method works for different distributions of response (which alters the distribution of ϕ), results from the interim analysis for different data sets can be found in rows three through five of Table 3.1 and in Figures 3.4–3.6. The narrow range for η in Table 3.1 will be addressed later in this sub-section. Furthermore, the table contains PP for $n_{\text{obs}} = 45, 60, \text{and } 75$ to compare how the choice of θ_L and θ_U can change decisions. For example, if $\theta_U = 0.99$ instead, Data sets 4 and 12 would no longer be able to stop testing after $n_{\text{obs}} = 45$. Additionally, Table 3.1 highlights that the number of observations required before a test can be stopped depends on the variability within the data. Data set 4 and 11 were generated from the same true model parameters, but the variance was higher for data set 11; data set 4 can be stopped early, but data set 11 requires seeing all observations before a conclusion can be made.

	Data Set		1	2	3	4	5	6	7
	MNP†	n_{obs}	η =343	η =345	η =347	η =349	η =351	η =353	η =355
PP	8	45	<0.0001‡	0.0412	0.6281	0.9668	>0.9999‡	>0.9999‡	>0.9999‡
	13	60	<0.0001‡	0.1116	0.8702	0.9990	>0.9999‡	>0.9999‡	>0.9999‡
	18	75	<0.0001‡	0.1781	0.9748	>0.9999‡	>0.9999‡	>0.9999‡	>0.9999‡
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		80	0.7856	0.7968	0.8059	0.8134	0.8212	0.8302	0.8380

(a) Error Transformation 1 (Smallest Variance)

	Data Set		8	9	10	11	12	13	14
	MNP†	n_{obs}	η =343	η =345	η =347	η =349	η =351	η =353	η =355
PP	8	45	<0.0001‡	<0.0001‡	0.0633	0.5634	0.9588	0.9980	>0.9999‡
	13	60	<0.0001‡	<0.0001‡	0.2968	0.7918	0.9960	>0.9999‡	>0.9999‡
	18	75	<0.0001‡	<0.0001‡	0.1861	0.9014	>0.9999‡	>0.9999‡	>0.9999‡
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		80	0.7812	0.7879	0.7959	0.8018	0.8096	0.8172	0.8248

(b) Error Transformation 2

	Data Set		15	16	17	18	19	20	21
	MNP†	n_{obs}	η =343	η =345	η =347	η =349	η =351	η =353	η =355
PP	8	45	<0.0001‡	<0.0001‡	0.0040	0.0996	0.5191	0.8874	0.9809
	13	60	<0.0001‡	<0.0001‡	0.0040	0.2103	0.7787	0.9849	>0.9999‡
	18	75	<0.0001‡	<0.0001‡	0.0050	0.3089	0.9336	>0.9999‡	>0.9999‡
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		80	0.7802	0.7885	0.7916	0.7998	0.8062	0.8123	0.8195

(c) Error Transformation 3 (Largest Variance)

* Definitions for Error Transformation 1, 2, and 3 can be found in Appendix B.

† MNP is the minimum number of observations informing the posterior of each model parameter.

‡ Computationally, this was numerically indistinguishable from 0 or 1, but not actually 0 or 1.

Table 3.1: PP and $\Pr_{\phi|X}(\phi \geq 400)$ for Various n_{obs}s and ηs

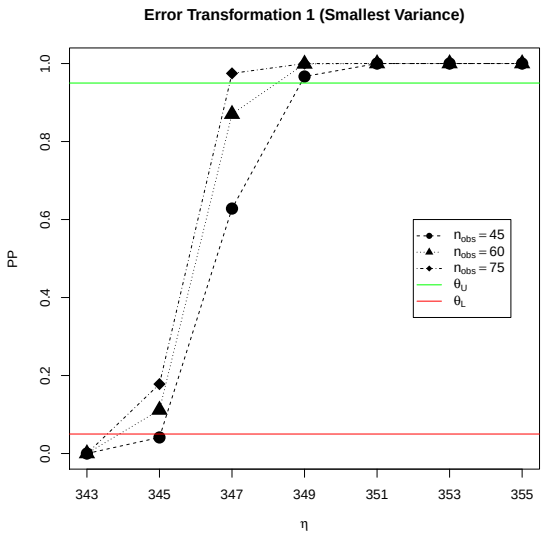

Figure 3.4: OT Data Sets 1–7

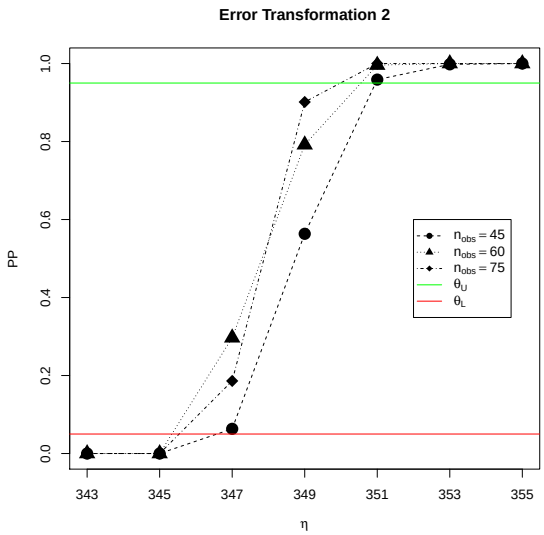

Figure 3.5: OT Data Sets 8–14

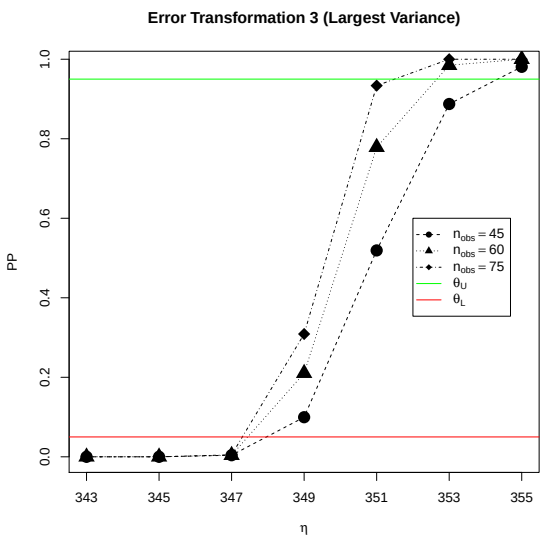

Figure 3.6: OT Data Sets 15–21

At $n_{\text{obs}} = 45$ (with $\theta_L = 0.05$ and $\theta_U = 0.95$), data sets 1, 2, 8, 9, 15, 16, and 17 demonstrate tests which could be stopped early with confidence that the electric semi-truck would fail to obtain a mean number of miles traveled on one charge greater than 400 miles at $\theta_T = 0.8$. Data sets 4, 5, 6, 7, 12, 13, 14, and 21 demonstrate tests which could be stopped early with confidence that the electric semi-truck would obtain the threshold at θ_T. Data sets 3, 10, 11, 18, 19, and 20 demonstrate tests which require more testing at $n_{\text{obs}} = 45$.

Data sets 10, 11, 18, and 19 demonstrate tests which would require accomplishing all test events before evaluating the measure. These data sets would require calculating $\Pr_{\phi|X}(\phi \geq 400)$ and comparing that to $\theta_T = 0.8$ at the end of the test to evaluate the measure. This would be done as detailed in Section 2.3.2, and the results of Section 2.3.2 can be seen in the sixth row of Table 3.1.

It is worth noting that using PP (in this example) results in decisions to stop a test early, except for when the posterior probability is very close to 0.8—that is to say, when the $(1-\theta)^{\text{th}}$ quantile for the distribution of ϕ is very close to ϕ_0. This narrow range of η values was chosen because slightly larger or smaller values of η resulted in PPs that were all <0.0001 or >0.9999. Outside of a narrow range for η, the method provides conclusive results in our simulation; it is only in this narrow range that PPs may determine more testing is required.

Chapter 4

Developing Informative Priors from Developmental Testing

Chapter 2 presented a method for transitioning the current operational testing (OT) paradigm into a Bayesian framework; Chapter 3 leveraged this method to create a dynamic operational test, allowing for the possibility to stop OT early and save resources. Chapter 4 presents a method that allows for further resources to be saved, extending Chapters 2 and 3, by proposing a novel method for incorporating developmental testing (DT) information into OT. A strength of the Bayesian paradigm is that it allows for the explicit use of all available information—to include subject matter expert (SME) opinion. While frequentists are constrained to only including data in an analysis (that is to say, only including things that can be observed), Bayesians can easily consider both data and SME opinion, or any other related information that could be constructed. This can be accomplished through the development and use of priors (D. A. Berry 1993; S. M. Berry et al. 2011; R. Christensen et al. 2011).

Section 4.1 is a review of priors that relate to this research, and Section 4.2 proposes a novel prior for incorporating DT data into OT. Finally, Section 4.3 im-

plements this prior using the electric semi-truck example, and examines how the prior influences conclusions about a measure when using predictive probability (PP) and posterior probability. For a method to be effective, the interim analysis must be accomplished in a timely manner so that decision makers can make an informed decision about stopping a test without prolonging OT. Therefore, an overarching theme of this chapter is that DT information should be incorporated into OT via a computationally efficient method—one that can be used multiple times over a short period of time to support a decision (such as employing interim analysis in OT to inform a decision to stop testing or not).

4.1 Priors

From a Bayesian perspective, every parameter in a data model is unknown and requires an associated prior distribution. Prior distributions represent an individual's beliefs about model parameters, quantifying the uncertainty surrounding those parameters. As R. Christensen et al. (2011, hereafter referred to as BIDA) discuss, prior distributions do not necessarily exhibit the true nature of parameters; rather, they exhibit an individual's understanding of those parameters. As such, priors are not perfect representations of nature (nor do they need to be perfect representations of an individual's beliefs). Instead, priors are a means of incorporating key information about model parameters into a statistical analysis (BIDA). Examples include the relative information included in a prior (discussed in Sections 4.1.1–4.1.6) and the support of model parameters. The posterior distribution for model parameters cannot have support outside the support of the priors; therefore, it is imperative that prior distributions have a reasonable support. The support of the priors should also mimic the support of model parameters as much as possible. If the support for the priors is too narrow, the posterior distribution may not capture all relevant features

of the model parameters. On the other hand, if the support for the priors is too wide, it can lead to computational inefficiencies and take longer to obtain the stationarity required for conducting appropriate inferences.

As Bayesian inference depends on the posterior distribution, which depends on prior distributions, prior distributions affect Bayesian inferences (S. M. Berry et al. 2011, hereafter referred to as BAMCT). However, despite this, "two reasonably open-minded people will eventually come to agree if both are exposed to the same data and use Bayes' Theorem" (BAMCT). Moreover, "even investigators with wildly dissimilar prior beliefs can ultimately come to agreement once sufficient data have been accumulated" (BAMCT). As such, the influence a prior has on the posterior distribution tends to decrease as the sample size increases and overwhelms prior information. However, the cost of increasing the sample size must be weighed against the risk of making decisions that depend too heavily on prior information (BAMCT). Not only does this highlight the importance of sample sizes, but this also highlights the importance of sensitivity analysis—evaluating the effect that changing priors has on conclusions. Frequently, sensitivity analysis is accomplished by comparing results based on an informative prior (see Sections 4.1.3–4.1.6) to those based on a reference prior (see Sections 4.1.1–4.1.2).

Another key concept related to priors is *conjugate priors*, which are used throughout this research. A conjugate prior is obtained when, "... in terms of θ, the prior density $p(\theta)$ has the same functional form as the sampling density $f(y|\theta)$ so that, after applying Bayes' Theorem, the posterior has the same functional form as the prior" (BIDA). For example, a $\text{Bin}(n, \theta)$ data model's conjugate prior on θ is a beta distribution, resulting in a beta posterior distribution. When sampling methods use conjugate priors (such as in this research), the computational efficiency often increases relative to the use of non-conjugate priors. However, as BIDA discuss, the convenience of using conjugate priors must be counterbalanced against selecting a

prior that accurately reflects the prior beliefs about parameters.

The remainder of this section is dedicated to discussing the amount of information contained in different types of priors related to this research. It begins with priors that strive to provide no information (reference priors) and builds up to priors that contain available information (informative priors).

4.1.1 Reference Priors

Reference priors (also known as convenience priors or noninformative priors) attempt to incorporate no prior information. While it is never possible to incorporate no prior information, reference priors are easily overwhelmed by the data, and, therefore, they have minimal influence on the posterior distribution (BIDA). Three commonly used types of reference priors are flat priors, improper priors, and Jeffreys' priors (although these types of priors are not necessarily mutually exclusive priors). Flat priors are priors with the form $p(\theta) = k$, for some constant k. Improper priors are priors that do not have a closed and bounded support, integrating to infinity. Jeffreys' priors are proportional to the square root of the Fisher information (BIDA). "Bayesians use flat or otherwise improper [reference] priors in situations where prior knowledge is vague relative to the information in the likelihood, or in settings where we want the data (and not the prior) to dominate the determination of the posterior" (BAMCT). Reference priors, therefore, attempt to demonstrate a lack of knowledge surrounding model parameters.

While reference priors may be convenient, they can cause inferential issues. As Gelman et al. (2014, hereafter referred to as BDA) discuss, care needs to be taken when using improper priors, as they are not guaranteed to result in a proper posterior distribution that integrates to one. Additionally, an improper prior can place a majority of the prior probability in strange locations. For example, the Beta(0,0)

improper prior places most of the prior probability on values close to 0 and 1—something that does not reflect the desire to demonstrate little prior knowledge about a parameter that lives between 0 and 1. Furthermore, improper priors are an example of a computationally inefficient prior, as too wide of a support can allow the sampling method to wander through the support of the prior for a non-trivial amount of time before it arrives in the stationary posterior distribution. In addition to these issues, flat priors also look different under parameter transformation. Consider the univariate parameter, θ, whose support is the entire real line; a uniform flat prior on θ will not result in a uniform flat prior on θ^2. This is an issue, as the same amount of prior information should be reflected in both θ and θ^2. Therefore, any prior that strives to provide no information about a parameter should also provide no information about a transformation of that parameter (Raiffa and Sclaifer, 1961, as cited in BIDA). In contrast, Jeffreys' priors are invariant under transformation (BDA). However, when "[u]sing Jeffreys' priors, models with proportional likelihoods lead to different inferences (because they have different priors) so the likelihood principle is violated. Similarly, stopping rules can change statistical inferences so the stopping rule principle is violated" (BIDA). Given that the likelihood principle and stopping rule principle are foundational to Bayesian inference, this creates problems from a foundational perspective.

As addressed in Section 3.2, it is known that using reference priors can result in high stopping rates for the methods presented in this research, which may be undesirable for testing (Saville et al. 2014). Therefore, careful consideration must be given when employing the methods presented in this research and using reference priors. How reference priors are used in this research will be addressed in Section 4.2.

4.1.2 Weakly Informative Priors

Instead of reference priors that strive to provide no information about the parameters, BDA suggest priors that incorporate an understanding of the physical limitations of the parameters. A prior that is designed to consider bounds of the physical world ensures that the resulting posterior distribution is reasonable, while still remaining vague enough to use for convenience (BDA). Such a prior distribution is referred to as a weakly informative prior distribution. Weakly informative priors provide minimal information about parameters, while avoiding computational issues and resulting in a reasonable posterior distribution. BDA provide the following reasons for why weakly informative priors might be used: "to describe the model more conveniently; because it may be difficult to express knowledge accurately in probabilistic form; to simplify computations; or perhaps to avoid using a possibly unreliable source of information." Weakly informative priors demonstrate that there is always some information available about parameters—even if that information simply limits the range of values which a parameter can reasonably take on (BDA, BIDA). For example, the discussion for creating priors in Section 2.3.1 resulted in weakly informative priors. Consider the baseline parameter, η, for the reference cell model. A Normal(400, 100^2) prior on η was selected because it placed 95% of the probability between 200 and 600 miles. By focusing on predicting what η could look like in the extremes, priors were created that placed a majority (95%) of the probability on reasonable values η could take on, while still allowing for a small chance (5%) of extremely unlikely values. While BDA refers to these types of priors as weakly informative priors, BIDA refers to these types of priors as reference priors; this research adopts the BIDA nomenclature, referring to them as reference priors in practice as well.

Although weakly informative priors do not have some of the problems other reference priors have, they also do not make full use of the information available. As

BDA details, a weakly informative prior "is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available."

4.1.3 Power Priors

As discussed in Chapter 1, OT alone may not provide enough data to make conclusive decisions; in such circumstances, formally eliciting and incorporating prior information is an important advantage of the Bayesian framework (BAMCT). Weakly informative priors and reference priors, while convenient, do not provide the advantage of augmenting small data sets with other relevant information. When historical data from a previous similar experiment are available, the data can be used to create an informative prior for the current experiment. A traditional approach for directly incorporating historical data into a Bayesian analysis is to use the posterior distribution of the historical experiment as the prior for the current experiment (Duan, Ye, and Smith 2006). However, this approach assumes that the data from the historical experiment and the current experiment are exchangeable—an assumption that may be difficult to satisfy (Duan, Ye, and Smith 2006). In contrast to this approach, Ibrahim and M.-H. Chen (2000) proposed a power prior. In cases where the historical data are similar, but not fully commensurate, analysts may wish to account for this by down-weighting the influence of the historical data on posterior distribution (Neelon and O'Malley 2010). For example, data from a drug trial in New York may inform a trial of the same drug in San Francisco; however, population differences may drive a desire to down-weight the New York trial information. Similarly, operational testers may wish to down-weight incorporated DT data to account for differences between OT and DT (e.g., differences in testing environment or system employment). This is the key concept of power priors—the influence of historical data on posterior is controlled by raising the likelihood for the historical data to a power between 0 and 1. Power priors are beneficial in that they formalize a way for incorporating

potentially dissimilar historical data into the current experiment, thereby creating an informative prior that does not assume the historical and current data are exchangeable. As such, Ibrahim and M.-H. Chen (2000) describe how a power prior can be seen as a generalization of the usual Bayesian updating methods. Furthermore, it has been shown that power priors (as originally proposed) are an optimal class of informative priors—optimal in that they minimize the Kullback-Leibler (KL) divergence between the posterior resulting from not incorporating historical data and the posterior resulting from pooling the historical and current data (Ibrahim, M.-H. Chen, Gwon, et al. 2015; Ibrahim, M.-H. Chen, and Sinha 2003).

To form a power prior, let Y_0 be the set of responses from the historical experiment, and let X_0 be the corresponding design matrix; further, let $D_0 = (Y_0, X_0, n_0)$. The likelihood function for the historical data is $L(\boldsymbol{\theta}|D_0)$, where $\boldsymbol{\theta}$ is the set of model parameters. Similarly, let Y be the set of responses from the current experiment, X be the corresponding design matrix, and $D = (Y, X, n)$. The likelihood function for the current data is $L(\boldsymbol{\theta}|D)$. The power prior assumes that all parameters are the same between the historical and current data model (an assumption that will be relaxed later).

When there is historical data, D_0, from a previous experiment, the power prior is

$$p(\boldsymbol{\theta}|D_0, a_0) \propto (L(\boldsymbol{\theta}|D_0))^{a_0} p_0(\boldsymbol{\theta}),$$

"where $0 \leq a_0 \leq 1$ is a [fixed] scalar parameter and $p_0(\boldsymbol{\theta})$ is the *initial prior* for $\boldsymbol{\theta}$ before the historical data D_0 is observed" (Ibrahim, M.-H. Chen, Gwon, et al. 2015; Ibrahim and M.-H. Chen 2000). The fixed a_0 controls how influential the historical data is on the posterior distribution. To illustrate this, take $p_0(\boldsymbol{\theta}) \propto 1$ (a flat prior); a_0 would then drive the shape (informativeness) of the prior. The power prior would be a fully informative prior when $a_0 = 1$ (i.e., fully pooling historical and current data); a flat prior when $a_0 = 0$ (i.e., no historical information is incorporated); and something between a flat and fully informative prior when $0 < a_0 < 1$ (i.e., historical

information is down-weighted). "One of the main roles of a_0 is that it controls the heaviness of the tails of the prior for θ. As a_0 becomes smaller, the tails of [the prior] becomes heaver" (Ibrahim, M.-H. Chen, Gwon, et al. 2015).

The role of a_0 can be more easily understood through the following example: suppose the historical data are n_0 random observations (y_{0i}, $i = 1, \ldots, n_0$) from a Normal$(0, 0.138)$ distribution. Furthermore, suppose that the mean is unknown, the precision is known (where $\frac{1}{\tau} = \sigma^2$), and $p_0(\mu) \propto 1$. The power prior for this construct would be

$$\begin{aligned}
p(\mu|\tau, D_0, a_0) &\propto (L(\mu|\tau, D_0))^{a_0} p_0(\mu) \\
&\propto \left[\left(\frac{\tau}{2} \right)^{\frac{n_0}{2}} \exp\left(-\frac{\tau}{2} \sum_{i=1}^{n_0} (y_{0i} - \mu)^2 \right) \right]^{a_0} (1) \\
&= \left(\frac{\tau}{2} \right)^{\frac{n_0 a_0}{2}} \exp\left(-\frac{\tau a_0}{2} \sum_{i=1}^{n_0} (y_{0i} - \mu)^2 \right) \\
&= \left(\frac{\tau}{2} \right)^{\frac{n_0 a_0}{2}} \exp\left(-\frac{\tau a_0}{2} \left(n_0 (\bar{y}_0 - \mu)^2 + \sum_{i=1}^{n_0} (y_{0i} - \bar{y}_0)^2 \right) \right) \\
&\propto \exp\left(-\frac{\tau a_0 n_0}{2} (\bar{y}_0 - \mu)^2 \right)
\end{aligned}$$

Therefore, when $p_0(\mu) \propto 1$ and τ is known, $p(\mu|\tau, D_0, a_0) \sim \text{Normal}\big(\bar{y}_0, \frac{1}{\tau a_0 n_0}\big)$. As such, "$a_0$ can be interpreted as a precision parameter for the historical data" (Ibrahim, M.-H. Chen, Gwon, et al. 2015). Alternatively, "[b]ecause $0 \leq a_0 \leq 1$, we might also think of $a_0 n_0$ as the 'effective' number of historical controls being incorporated into our analysis" (Hobbs et al. 2011). While this interpretation of a_0 breaks down for other distributions (e.g. consider if τ was unknown—this would result in a normal-gamma distribution, where three of the four distribution parameters depended on a_0, multiplied by $2^{-(\frac{n_0 a_0}{2})}$), it providers the reader with some intuition for how a_0 is impacting the prior. This impact can be seen visually for this example in Figure 4.1, which demonstrates how the shape (informativeness) of $p(\mu|\tau, D_0, a_0)$ changes based on different values of a_0. Ultimately, $p(\mu|D_0, a_0) \to p_0(\mu)$ as $a_0 \to 0$,

which is a flat prior for this example.

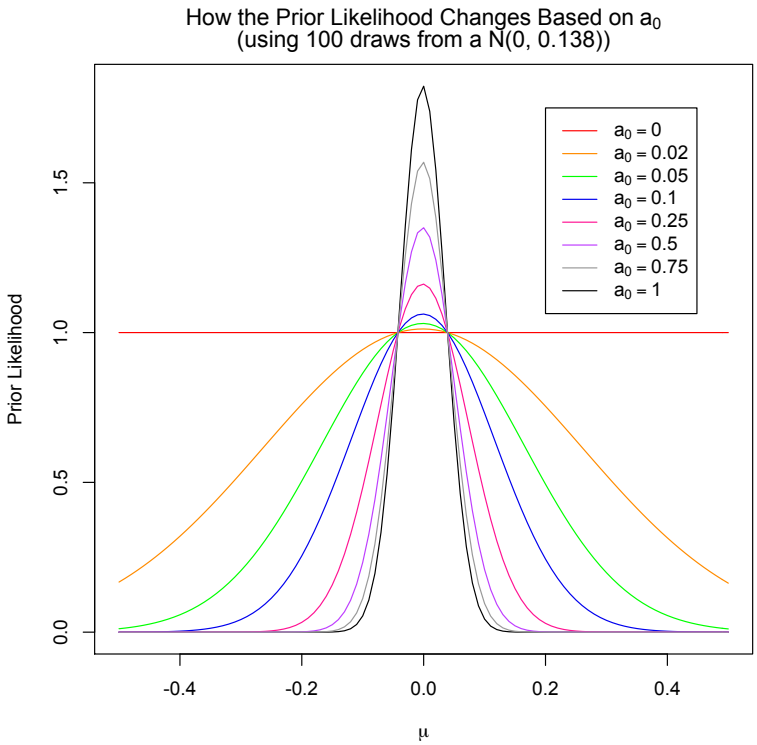

Figure 4.1: How the Shape of the Likelihood Changes for Normal(0,1) Data Based on a_0

When a_0 is fixed, methods such as eliciting subject matter expert (SME) opinion or model selection methods are recommended for selecting a_0, in conjunction with a sensitivity analysis on a_0 (Ibrahim, M.-H. Chen, and Sinha 2003; Ibrahim, M.-H. Chen, Gwon, et al. 2015). Additionally, a fixed a_0 leads to more computationally efficient MCMC sampling methods, such as a Gibbs sampler (Ibrahim, M.-H. Chen, Gwon, et al. 2015). While the sampling method itself maybe computationally efficient, the added burden of selecting a fixed a_0 and conducting sensitivity analysis on a_0 may make the entire process unsuitable for interim analysis within a short amount of time. Furthermore, the risk of selecting a_0 incorrectly (specifically, picking a_0 to be

closer to 1 than the data would suggest) can lead to dire outcomes: should a system be accepted that only met requirements by allowing dissimilar DT information to influence the posterior, the system may fail in the field, possibly to the detriment of the mission and lives. Therefore, in spite of the benefits of a fixed a_0, we prefer to consider a random a_0.

While the originators of the power prior and their coauthors continue to recommend that a_0 be taken as fixed (M.-H. Chen, Ibrahim, Amy Xia, et al. 2014; Ibrahim, M.-H. Chen, Gwon, et al. 2015), a joint power prior was also introduced by Ibrahim and M.-H. Chen (2000) which allowed for a random a_0:

$$p(\boldsymbol{\theta}, a_0|D_0) = (L(\boldsymbol{\theta}|D_0))^{a_0} p_0(\boldsymbol{\theta}) p_0(a_0),$$

where $p_0(a_0)$ is the initial prior on a_0. When the historical data and the current data are found to be dissimilar, a_0 creates a more diffuse prior to ensure the support is appropriate for the current data—thereby decreasing the influence that the historical data has on the posterior. Alternatively, when the historical data and the current data are found to be similar, a_0 does not need to create a diffuse prior to maintain the appropriate support—thereby allowing the historical data to have a larger influence on the posterior distribution. Therefore, a random a_0 can be seen as an indirect measure of the commensurability of the two data sets. Recall from Chapter 1 that Dewald et al. (2016) presented a method for creating priors based on summary statistics from previous simulations. While they did not use a power prior, their method also involved down-weighting the information in the prior by a specified weight; however, they suggested that, while more complex, the weight could have been selected based on the data by means such as a joint power prior.

One criticism of the joint power prior is that it violates the likelihood principle, since multiplying the likelihood function by a constant would change the joint prior (Duan, Ye, and Smith 2006). The next section introduces a variant of the joint power prior that avoids violating the likelihood principle, while maintaining a random a_0.

4.1.4 Normalized Power Priors

In 2006, Duan, Ye, and Smith proposed the normalized power prior (NPP), which has the following form:

$$p(\boldsymbol{\theta}, a_0|D_0) = C\left[\frac{(L(\boldsymbol{\theta}|D_0))^{a_0}p_0(\boldsymbol{\theta})}{\int(L(\boldsymbol{\theta}|D_0))^{a_0}p_0(\boldsymbol{\theta})d\boldsymbol{\theta}}\right]p_0(a_0)I_A(a_0), \tag{4.1}$$

where

$$I_A(a_0) = \begin{cases} 1, & \text{if } a_0 \in A \\ 0, & \text{otherwise} \end{cases}$$

As in the joint power prior, $p_0(\boldsymbol{\theta})$ and $p_0(a_0)$ are initial priors. Additionally, C is a normalizing constant and A is the region of a_0 such that:

$$A = \left\{a_0 : 0 < \int_\Theta (L(\boldsymbol{\theta}|D_0)^{a_0})p_0(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty\right\}.$$

While the initial assumption of a NPP is that $0 \leq a_0 \leq 1$, the region A may further restrict a_0 (as will be seen in Section 4.6.7).

One criticism of NPP is that the "formulation is even more computationally extensive than the joint power prior formulation for models other than normal linear regression models since for most non-normal models, an analytical evaluation of the integral [in equation (4.1)] is not available, which poses a huge challenge in sampling from the resulting posterior distribution and computing the posterior quantities of interest" (Ibrahim, M.-H. Chen, Gwon, et al. 2015). Furthermore, even when the integral in equation (4.1) is tractable, the full conditionals are not guaranteed to be recognizable distributions. Such cases would require a sampling method such as Metropolis-within-Gibbs, which is less efficient than a Gibbs sampler. The potential for NPP to be computationally inefficient will be re-considered in Section 4.2.

Another criticism of NPP is that it may overly down-weight historical data when the current and historical data are not approximately the same (Ibrahim, M.-H.

Chen, Gwon, et al. 2015; Neelon and O'Malley 2010). However, for an appropriate study goal, this does not invalidate NPP. If the goal of the study is to only incorporate historical data when it is approximately the same as the current data, NPP can be a reasonable choice for a prior (Neelon and O'Malley 2010). Connected with this criticism is the informativeness of the prior on a_0; because of the tendency to overly down-weight historical information, a fairly informative prior on a_0 is required to bound a_0 away from 0 and allow for sufficient borrowing (Neelon and O'Malley 2010; Hobbs et al. 2011). "In fact, under a flat Beta$(1, 1)$ prior on $[a_0]$, the marginal posterior for $[a_0]$ is flat for two identical [historical and current] datasets regardless of the sample sizes" (Hobbs et al. 2011). This criticism will be re-considered in Section 4.3.

While NPP incorporates a random a_0 without violating the likelihood principle, an assumption of NPP, joint power priors, and power priors is that all model parameters are the same between the historical and current data model. This is an unsuitably restrictive assumption that would make NPP difficult to use in practice for incorporating DT information into OT.

4.1.5 Partial Borrowing Power Priors

Another variant of the power prior is the partial borrowing power prior (PBPP). PBPP not only allows for the historical data model to be a subset of the current data model parameters, but also allows for additional model parameters (such as nuisance parameters or covariates) unique to the historical data model and current data model (Ibrahim, M.-H. Chen, Xia, et al. 2012; Ibrahim, M.-H. Chen, Gwon, et al. 2015). A special case of PBPP is when the parameters in the historical data model are a strict subset of the parameters in the current data model, as is assumed in this research. For more information about the general case, see Ibrahim, M.-H.

Chen, Xia, et al. 2012, Ibrahim, M.-H. Chen, Gwon, et al. 2015, and Psioda and Ibrahim 2019.

Before presenting the special case of PBPP (hereafter referred to as PBPP for simplicity), additional notation must be defined. The set of model parameters, $\boldsymbol{\theta}$, is partitioned into two pieces: the set of model parameters that are common to the historical and current data models ($\boldsymbol{\theta}_0$), and the set of model parameters that are only in the current data model ($\boldsymbol{\theta}_1$). Therefore, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. The PBPP, as first proposed, is an extension of the power prior and has the following form:

$$\begin{aligned} p(\boldsymbol{\theta}|a_0, D_0) &= p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|a_0, D_0) \\ &\propto (L(\boldsymbol{\theta}_0|D_0))^{a_0} p_0(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \end{aligned}$$

(Ibrahim, M.-H. Chen, Xia, et al. 2012; Psioda and Ibrahim 2019). When $p_0(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = p_0(\boldsymbol{\theta}_0)p_0(\boldsymbol{\theta}_1)$, the PBPP can be rewritten as

$$\begin{aligned} p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|a_0, D_0) &\propto (L(\boldsymbol{\theta}_0|D_0))^{a_0} p_0(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \\ &= (L(\boldsymbol{\theta}_0|D_0))^{a_0} p_0(\boldsymbol{\theta}_0)p_0(\boldsymbol{\theta}_1) \\ &= p(\boldsymbol{\theta}_0|a_0, D_0)p(\boldsymbol{\theta}_1), \end{aligned} \tag{4.2}$$

where $p(\boldsymbol{\theta}_0|a_0, D_0)$ is a power prior and $p(\boldsymbol{\theta}_1)$ is an independent prior.

It is worth noting that the DT model parameters being a subset of OT model parameters is not the only construct that could be considered. Hobbs et al. (2011) describe how NPP does not *directly* measure the commensurability between the historical and the current data. Therefore, they proposed commensurate power priors, which assume two different parameters in the historical data and the current data. While such a prior may be reasonable, it would require an extensive understanding of how to model DT information, which is beyond the scope of this research. While PBPP is still restrictive (and a simplistic representation of DT), PBPP is more flexible than NPP. After establishing a method under this simplifying assumption, future work will explore more complex modeling assumptions. Although less

restrictive than NPP, PBPP incorporates a fixed a_0; the next section extends PBPP to incorporate a random a_0.

4.1.6 Normalized Partial Borrowing Power Prior

In 2015, Ibrahim, M.-H. Chen, Gwon, et al. stated that PBPP could be adapted in many ways—to include using NPP instead of a power prior in equation (4.2). Using NPP would extend the original PBPP to a PBPP with a random a_0 that does not violate the likelihood principle. Under this construct, the power prior in equation (4.2) would be changed to the NPP from equation (4.1), resulting in

$$\begin{aligned} p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, a_0 | D_0) &\propto p(\boldsymbol{\theta}_0, a_0 | D_0) p(\boldsymbol{\theta}_1) \\ &= C \left[\frac{(L(\boldsymbol{\theta}_0 | D_0))^{a_0} p_0(\boldsymbol{\theta}_0)}{\int (L(\boldsymbol{\theta}_0 | D_0))^{a_0} p_0(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0} \right] p_0(a_0) I_A(a_0) p(\boldsymbol{\theta}_1). \end{aligned} \tag{4.3}$$

This research refers to a PBPP that uses NPP as a normalized partial borrowing power prior (NPBPP). The NPBPP in equation (4.3) has only been used in literature on one occasion: M.-H. Chen, Ibrahim, Lam, et al. (2011). In a medical device trial that was comparing a control device to a new device, M.-H. Chen, Ibrahim, Lam, et al. (2011) used NPBPP to only borrow historical information for the control device, as no historical data for the new device was available.

As will be shown in (and to compare with) Section 4.2, a prior based on conditional probability could be used instead of $p_0(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = p_0(\boldsymbol{\theta}_0) p_0(\boldsymbol{\theta}_1)$ in equation (4.2):

$$\begin{aligned} p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, a_0 | D_0) &\propto (L(\boldsymbol{\theta}_0 | D_0))^{a_0} p_0(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) p_0(a_0) \\ &= (L(\boldsymbol{\theta}_0 | D_0))^{a_0} p_0(\boldsymbol{\theta}_0) p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0) p_0(a_0) \\ &= p(\boldsymbol{\theta}_0, a_0 | D_0) p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0), \end{aligned} \tag{4.4}$$

where $p(\boldsymbol{\theta}_0, a_0|D_0)$ is a NPP. This variation has not been used in literature; however, it can ensure a certain amount of conjugacy in some cases, as will be seen in Section 4.1.7.

Recall that the current OT construct qualifies non-OT data for use if the data meets certain conditions—conditions that ensure non-OT data are exchangeable with OT data. These restrictive conditions can result in disregarding beneficial functionally representative non-OT data. Using NPBPP aligns with the intent of the current OT construct for qualifying non-OT data, without being as restrictive—functionally representative DT data can be incorporated into NPBPP and down-weighted by a_0 depending on the degree of similarity between the two data sets. We define functionally representative data to be data from a meaningfully similar and mature system in DT that provides information about OT model parameters, but that was collected in an environment that was not fully operationally representative. By only considering a mature system that is meaningfully the same as OT as it relates to a given measure, the risk that DT data could be disinformative for OT is reduced. Furthermore, given that the system is meaningfully the same, we assume that the variance parameter $(\frac{1}{\tau})$ is the same in both DT and OT. This assumption does not imply that the spread of the OT responses compared to the spread of the DT responses will be the same; rather, it implies that the errors around the fitted values have equal variance in both OT and DT. While it can still be argued that DT has a smaller variance than OT, we contend that the apparently smaller variance is artificially created by the more controlled DT environment. It appears smaller, but it is not truly smaller because (by assumption) the system should be meaningfully the same in DT as in OT. This assumption is further examined in Section 4.3. Furthermore, recall from Chapter 1 that OT and DT may be interested in different requirements. It is worth noting that the NPBPP construct does not assume that the measure being evaluated (or the requirement that measure is derived from) in OT is the same as in DT. The NPBPP construct only assumes that the response variable is the same in OT and DT.

4.1.7 Normalized Partial Borrowing Power Prior - Normal Linear Regression Model

Finally, we give a summary of what this looks like for a normal linear regression model, which relates to the construct for the electric semi-truck example. As highlighted in Section 4.1.4, even when attention is restricted to normal linear regression models with tractable integrals in equation (4.1), NPP (and, by extension NPBPP) may result in the need for a computationally inefficient sampling method. This section will shown that, under a DT / OT construct, using NPBPP will result in the need for a computationally inefficient Metropolis-within-Gibbs sampler. First, the previously introduced notation is adjusted to reflect an ANOVA model (using a reference cell model set-up) for a DT / OT construct in generality. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \tau)$, where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0p_0})'$ are the p_0 model parameters that are common to both the DT and the OT data models. Similarly, let $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p_1})'$ are the p_1 model parameters that are only in OT data model. The likelihood function for the DT data (D_0) is then $L(\boldsymbol{\beta}_0, \tau | D_0)$; the likelihood function for the OT data (D) is $L(\boldsymbol{\beta}, \tau | D)$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)$. Having partitioned $\boldsymbol{\beta}$ into two pieces, the design matrix for OT, X, can be partitioned as well. X can be rewritten as $X = [X_1, X_2]$, where the columns of X_1 correspond to parameters that are in $\boldsymbol{\beta}_1$ and columns of X_2 correspond to parameters that are in $\boldsymbol{\beta}_0$. It is important to note that the entries in X_2 are the values from OT, not DT. With this partition, $X\boldsymbol{\beta}$ can be rewritten as follows:

$$X\boldsymbol{\beta} = [X_1, X_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_0 \end{bmatrix} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_0$$

Finally, X_0 is the design matrix for DT.

Following Duan, Ye, and Smith (2006), it is assumed that, given $\boldsymbol{\beta}$, D_0 and D are independent random samples from normal distributions. The following multivariate

normal data models can then be used for OT and DT, respectively:

$$D|\boldsymbol{\beta}, \tau \sim \mathcal{N}(X\boldsymbol{\beta}, \tau^{-1} I_n)$$
$$D_0|\boldsymbol{\beta}_0, \tau \sim \mathcal{N}(X_0\boldsymbol{\beta}_0, \tau^{-1} I_{n_0})$$

where I_{n_0} is a $n_0 \times n_0$ identity matrix and I_n is a $n \times n$ identity matrix. Then,

$$\begin{aligned} &L(\boldsymbol{\beta},\tau|D) \\ &\quad = (2\pi)^{-\frac{n}{2}} (\det(\tau^{-1} I_n))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left((X\boldsymbol{\beta} - Y)'(\tau^{-1} I_n)^{-1}(X\boldsymbol{\beta} - Y) \right) \right) \end{aligned}$$

For a $n \times n$ matrix, A, and a scalar value, c, the $\det(cA) = c^n \det(A)$:

$$\begin{aligned} &= (2\pi)^{-\frac{n}{2}} \left(\tau^{-n} \det(I_n) \right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left((X\boldsymbol{\beta} - Y)'(\tau^{-1} I_n)^{-1}(X\boldsymbol{\beta} - Y) \right) \right) \\ &= (2\pi)^{-\frac{n}{2}} (\tau)^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y) \right) \right) \end{aligned}$$

Similarly, $L(\boldsymbol{\beta}_0, \tau|D_0)$ is

$$L(\boldsymbol{\beta}_0, \tau|D_0) = (2\pi)^{-\frac{n_0}{2}} (\tau)^{\frac{n_0}{2}} \exp\left(-\frac{\tau}{2}\left((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) \right) \right)$$

Finally, $(L(\boldsymbol{\beta}_0, \tau|D_0))^{a_0}$ is

$$(L(\boldsymbol{\beta}_0, \tau|D_0))^{a_0} = (2\pi)^{-\frac{n_0 a_0}{2}} (\tau)^{\frac{n_0 a_0}{2}} \exp\left(-\frac{a_0\tau}{2}\left((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) \right) \right)$$

For consistency and comparison purposes, the same initial priors are used here as are used in Section 4.2 (where more detail for why these initial priors were selected can be found):

- $\boldsymbol{\beta}_1|\tau \sim \mathcal{N}(\boldsymbol{\mu}_1, \tau^{-1}\boldsymbol{\Lambda}_1^{-1})$ for some positive definite $\boldsymbol{\Lambda}_1$
- $p_0(\boldsymbol{\beta}_0) \propto 1$
- $\tau \sim \text{Gam}(\alpha_0, \gamma_0)$, using the shape / rate parameterization

- $a_0 \sim \text{Beta}(1,1)$

Note that these initial priors assume $p_0(\boldsymbol{\beta}_0, \tau) = p_0(\boldsymbol{\beta}_0)p_0(\tau)$. With this set-up, it can be shown that the NPBPP from equation (4.4) is

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, a_0 | D_0) \\
&\qquad = p(\boldsymbol{\beta}_0, \tau, a_0 | D_0) p_0(\boldsymbol{\beta}_1 | \tau) \\
&\qquad = C\left[\frac{(L(\boldsymbol{\beta}_0, \tau | D_0))^{a_0} p_0(\boldsymbol{\beta}_0) p_0(\tau)}{\int\int (L(\boldsymbol{\beta}_0, \tau | D_0))^{a_0} p_0(\boldsymbol{\beta}_0) p_0(\tau) d\boldsymbol{\beta}_0 d\tau}\right] p_0(a_0) I_A(a_0) p_0(\boldsymbol{\beta}_1 | \tau) \\
&\qquad = C(2\pi)^{-\frac{p_0}{2}} (\det((X_0' X_0)^{-1}))^{-\frac{1}{2}} (a_0)^{\frac{p_0}{2}} \frac{\gamma_{0n}^{\alpha_{0n}}}{\Gamma(\alpha_{0n})} (\tau)^{(\frac{n_0 a_0}{2} + \alpha_0) - 1} \\
&\qquad\qquad \times \exp\left(-\frac{a_0 \tau}{2} \left((X_0 \boldsymbol{\beta}_0 - Y_0)' (X_0 \boldsymbol{\beta}_0 - Y_0) \right) \right) \\
&\qquad\qquad \times \exp(-\gamma_0 \tau) (2\pi)^{-\frac{p_1}{2}} (\tau)^{\frac{p_1}{2}} (\det(\boldsymbol{\Lambda}_1^{-1}))^{-\frac{1}{2}} \\
&\qquad\qquad \times \exp\left(-\frac{1}{2} (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)' (\tau^{-1} \boldsymbol{\Lambda}_1^{-1})^{-1} (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1) \right) I_A(a_0),
\end{aligned}
$$

where

$$
\begin{aligned}
A &= \begin{cases} \left[\frac{p_0 - 2\alpha_0}{n_0}, 1\right] & \text{for } \alpha_0 < \frac{p_0}{2} \\ [0,1] & \text{for } \alpha_0 \geq \frac{p_0}{2} \end{cases} \\
\alpha_{0n} &= \frac{n_0 a_0 - p_0}{2} + \alpha_0 \\
\gamma_{0n} &= \frac{a_0}{2} (M_0 Y_0 - Y_0)' (M_0 Y_0 - Y_0) + \gamma_0
\end{aligned}
$$

Therefore, when $\alpha_0 < \frac{p_0}{2}$, a_0 will borrow at least some information from the historical data because the lower bound for a_0 will no longer be 0. Using this NPBPP, the posterior is:

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1, \tau, a_0|D_0, D) \\
&\quad \propto L(\boldsymbol{\beta}, \tau|D)p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, a_0|D_0) \\
&\quad = (2\pi)^{-\frac{n}{2}}(\tau)^{\frac{n}{2}} \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)\Big)\bigg) C(2\pi)^{-\frac{p_0}{2}} \\
&\qquad\qquad \times (\det((X_0'X_0)^{-1}))^{-\frac{1}{2}}(a_0)^{\frac{p_0}{2}} \frac{\gamma_{0n}^{\alpha_{0n}}}{\Gamma(\alpha_{0n})}(\tau)^{(\frac{n_0 a_0}{2}+\alpha_0)-1} \\
&\qquad\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0)\Big)\bigg) \exp(-\gamma_0\tau) \\
&\qquad\qquad \times (2\pi)^{-\frac{p_1}{2}}(\tau)^{\frac{p_1}{2}}(\det(\boldsymbol{\Lambda}_1^{-1}))^{-\frac{1}{2}} \\
&\qquad\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg) I_A(a_0).
\end{aligned}
$$

This posterior distribution does not result in a recognizable distribution; instead, full conditionals can be used to sample from the posterior. First, the following parameters are defined:

$$
\begin{aligned}
\boldsymbol{\Lambda}_{0f} &= (X_2'X_2) + a_0(X_0'X_0) \\
\boldsymbol{\mu}_{0f} &= \boldsymbol{\Lambda}_{0f}^{-1}\Big(X_2'Y - X_2'X_1\boldsymbol{\beta}_1 + a_0X_0'Y_0\Big) \\
\boldsymbol{\Lambda}_{1f} &= (X_1'X_1) + \boldsymbol{\Lambda}_1 \\
\boldsymbol{\mu}_{1f} &= \boldsymbol{\Lambda}_{1f}^{-1}\Big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big) \\
\gamma_f &= \frac{1}{2}\bigg[(X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y) + a_0(X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) \\
&\qquad\qquad + (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg] + \gamma_0 \\
\alpha_f &= \frac{n + n_0a_0 + p_1}{2} + \alpha_0.
\end{aligned}
$$

The full conditionals (see Appendix C for the derivations) for this posterior distribution are:

$$\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \tau, a_0, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{0f}, \tau^{-1}\boldsymbol{\Lambda}_{0f}^{-1})$$

$$\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_0, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{1f}, \tau^{-1}\boldsymbol{\Lambda}_{1f}^{-1})$$
$$\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D \sim \text{Gam}(\alpha_f, \gamma_f),$$

and

$$p(a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D)$$
$$\propto a_0^{\frac{p_0}{2}} \frac{\gamma_{0n}^{\alpha_{0n}}}{\Gamma(\alpha_{0n})}(\tau)^{(\frac{n_0 a_0}{2}+\alpha_0)-1} I_A(a_0) \exp\left(-\frac{a_0\tau}{2}(X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0)\right)$$

The full conditional for a_0 will not result in a recognizable distribution, which can be seen when considering the term $\gamma_{0n}^{\alpha_{0n}}$. Furthermore, when both $\boldsymbol{\beta}_0$ and τ are random, the full conditional for a_0 will not be recognizable, even if other initial priors are considered. Therefore, a computationally inefficient Metropolis-within-Gibbs sampler would be required.

4.2 A Novel Approach Based the Partial Borrowing Power Prior

When conducting interim analysis in an operational test, any computational inefficiencies can preclude a method from being suitable. To address this issue, a new prior is proposed that is both computationally efficient and maintains a random $\boldsymbol{\beta}_0$ and τ in the DT and OT data models.

4.2.1 The Conditional Normalized Partial Borrowing Power Prior

Recall, from equation (4.4), that NPBPP has the form

$$p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, a_0|D_0) \propto p(\boldsymbol{\theta}_0, a_0|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0),$$

where $p(\boldsymbol{\theta}_0, a_0|D_0)$ is a NPP. An alternative way to look at this equation is to consider $p(\boldsymbol{\theta}_0, a_0|D_0)$ a joint prior on $(\boldsymbol{\theta}_0, a_0)$, which happens to be a NPP. However, this is not the only way to construct a prior on $(\boldsymbol{\theta}_0, a_0)$. To illustrate this, consider a Normal$(\mu, \frac{1}{\tau})$, data model, where both μ and τ are unknown. One option for a prior is a joint prior on (μ, τ), such as the reference prior $p(\mu, \tau) = \frac{1}{\tau}$ (BIDA). Another option is a set of conjugate priors, such as

$$\begin{aligned}\mu|\tau &\sim \mathrm{N}\left(\mu_0, \frac{1}{\omega_0\tau}\right)\\ \tau &\sim \mathrm{Gam}\left(\frac{a}{2}, \frac{b}{2}\right)\end{aligned}$$

(BIDA). This set of conjugate priors is a result of conditional probability, which states that $p(\mu, \tau) = p(\mu|\tau)p(\tau)$. Returning to equation (4.4), instead of a joint prior on $(\boldsymbol{\theta}_0, a_0)$, a prior based on conditional probability could be used. By partitioning $\boldsymbol{\theta}_0$ into $(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02})$, where the prior on $\boldsymbol{\theta}_{01}$ is conditioned on $\boldsymbol{\theta}_{02}$, equation (4.4) can then be re-written as:

$$\begin{aligned}p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, a_0|D_0) &\propto p(\boldsymbol{\theta}_0, a_0|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)\\ &= p(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, a_0|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)\\ &= p(\boldsymbol{\theta}_{01}, a_0|\boldsymbol{\theta}_{02}, D_0)p(\boldsymbol{\theta}_{02}|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0). \qquad (4.5)\end{aligned}$$

This research proposes the following priors for equation (4.5): $p(\boldsymbol{\theta}_{01}, a_0|\boldsymbol{\theta}_{02}, D_0)$ is a NPP, $p(\boldsymbol{\theta}_{02}|D_0)$ is an appropriate prior on $\boldsymbol{\theta}_{02}$, and $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)$ remains the conditional prior on $\boldsymbol{\theta}_1$. This proposed prior is referred to as a conditional normalized partial borrowing power prior (CNPBPP), and has the following form:

$$\begin{aligned}&p(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1, a_0|D_0)\\ &\quad\propto p(\boldsymbol{\theta}_{01}, a_0|\boldsymbol{\theta}_{02}, D_0)p(\boldsymbol{\theta}_{02}|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)\\ &\quad= C\left[\frac{(L(\boldsymbol{\theta}_{01}|\boldsymbol{\theta}_{02}, D_0))^{a_0}p_0(\boldsymbol{\theta}_{01}|\boldsymbol{\theta}_{02})}{\int(L(\boldsymbol{\theta}_{01}|\boldsymbol{\theta}_{02}, D_0))^{a_0}p_0(\boldsymbol{\theta}_{01}|\boldsymbol{\theta}_{02})d\boldsymbol{\theta}_{01}}\right]p_0(a_0)I_A(a_0)p(\boldsymbol{\theta}_{02}|D_0)p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)\end{aligned}$$

(4.6)

Using the same set-up as in Section 4.1.7, the following will demonstrate that using CNPBPP can be more computationally efficient than using NPBPP. As in Section 4.1.7, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \tau)$, where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0p_0})'$; and let $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p_1})'$. The likelihood function for the DT data is $L(\boldsymbol{\beta}_0, \tau | D_0)$ and for the OT data is $L(\boldsymbol{\beta}, \tau | D)$. Let X_0 be the DT design matrix, and let X be the OT design matrix; X can be rewritten as $X = [X_1, X_2]$, where the columns of X_1 correspond to parameters that are in $\boldsymbol{\beta}_1$ and columns of X_2 correspond to parameters that are in $\boldsymbol{\beta}_0$. Finally, let $\boldsymbol{\theta}_{01} = \boldsymbol{\beta}_0$, and $\boldsymbol{\theta}_{02} = \tau$. Given $\boldsymbol{\beta}$, D_0 and D are assumed to be independent random samples from normal distributions:

$$D|\boldsymbol{\beta}, \tau \sim \mathcal{N}(X\boldsymbol{\beta}, \tau^{-1} I_n)$$
$$D_0|\boldsymbol{\beta}_0, \tau \sim \mathcal{N}(X_0\boldsymbol{\beta}_0, \tau^{-1} I_{n_0})$$

where I_{n_0} is a $n_0 \times n_0$ identity matrix and I_n is a $n \times n$ identity matrix.

Using the same initial priors as in Section 4.1.7,

- $\boldsymbol{\beta}_1|\tau \sim \mathcal{N}(\boldsymbol{\mu}_1, \tau^{-1}\boldsymbol{\Lambda}_1^{-1})$ for some positive definite $\boldsymbol{\Lambda}_1$
- $p_0(\boldsymbol{\beta}_0|\tau) \propto 1$
- $\tau \sim \text{Gam}(\alpha_0, \gamma_0)$, using the shape / rate parameterization
- $a_0 \sim \text{Beta}(1, 1)$

The prior on $\boldsymbol{\beta}_1$ and the prior on a_0 were selected to ensure conjugacy for the set-up presented here. The prior on $\boldsymbol{\beta}_0|\tau$ was taken to be a flat prior, as is done in many applications for a power prior (Ibrahim, M.-H. Chen, Gwon, et al. 2015). As highlighted in Section 4.1.1, reference priors such as this can be problematic for Bayesian inference; however, the power prior construct updates the initial prior for $\boldsymbol{\beta}_0$ with historical data to create an informative prior. Therefore, while the initial prior

on $\boldsymbol{\beta}_0$ is a reference prior, the final prior on $\boldsymbol{\beta}_0$ will be an informative prior (when $a_0 \neq 0$), thereby alleviating the problems that can arise when reference priors are used. Furthermore, this research assumes $p(\tau|D_0) = p(\tau)$ (in equation (4.5), $p(\boldsymbol{\theta}_{02}|D_0) = p(\boldsymbol{\theta}_{02})$)—a prior that balances a representation of prior beliefs with convenience. Finally, as in Chapters 2 and 3, it is assumed that the priors on the model parameters are independent; therefore, $p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau) = p(\boldsymbol{\beta}_1|\tau)$. These assumptions were also implicit in Section 4.1.7, for consistency and comparison purposes. While the prior on $\boldsymbol{\beta}_0$ and τ from Section 4.1.7 was written as "$p_0(\boldsymbol{\beta}_0, \tau) = p_0(\boldsymbol{\beta}_0)p_0(\tau)$", note that the $p_0(\boldsymbol{\beta}_0, \tau) = p_0(\boldsymbol{\beta}_0|\tau)p_0(\tau)$ construct used here will result in the same expression of prior beliefs about $\boldsymbol{\beta}_0$, since $p_0(\boldsymbol{\beta}_0|\tau) \propto 1$ does not depend on τ.

With this set-up, it can be shown that CNPBPP from equation (4.6) is

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1, \tau, a_0|D_0) \\
&\quad = C\left[\frac{(L(\boldsymbol{\beta}_0|\tau, D_0))^{a_0} p_0(\boldsymbol{\beta}_0|\tau)}{\int (L(\boldsymbol{\beta}_0|\tau, D_0))^{a_0} p_0(\boldsymbol{\beta}_0|\tau) d\boldsymbol{\beta}_0}\right] p_0(a_0) I_A(a_0) p(\tau|D_0) p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau) \\
&\quad = C\left[\frac{(L(\boldsymbol{\beta}_0|\tau, D_0))^{a_0} p_0(\boldsymbol{\beta}_0|\tau)}{\int (L(\boldsymbol{\beta}_0|\tau, D_0))^{a_0} p_0(\boldsymbol{\beta}_0|\tau) d\boldsymbol{\beta}_0}\right] p_0(a_0) I_A(a_0) p(\tau) p(\boldsymbol{\beta}_1|\tau) \\
&\quad \propto C(2\pi)^{-\frac{p_0}{2}} (\tau)^{\frac{p_0}{2}} a_0^{\frac{p_0}{2}} (\det((X_0'X_0)^{-1}))^{-\frac{1}{2}} \\
&\qquad\qquad \times \exp\left(-\frac{a_0\tau}{2}\left((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\right)\right) I_A(a_0) \\
&\qquad\qquad \times (2\pi)^{-\frac{p_1}{2}} (\det(\tau^{-1}\boldsymbol{\Lambda}_1^{-1}))^{-\frac{1}{2}} \\
&\qquad\qquad \times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\right) \\
&\qquad\qquad \times \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} (\tau)^{\alpha_0 - 1} \exp(-\gamma_0\tau),
\end{aligned}
$$

where $A = [0, 1]$. This allows a_0 to control the borrowing of information from DT as much or as little as the data suggests, regardless of the hyperparameters selected for the prior on τ; this is in contrast to NPBPP, whose region A depends on the

hyperparameter α_0. The posterior is then

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1, \tau, a_0 | D_0, D) \\
&\quad \propto L(\boldsymbol{\beta}, a_0 | D) p(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, a_0 | D_0) \\
&\quad \propto (2\pi)^{-\frac{n}{2}} (\tau)^{\frac{n}{2}} \exp\bigg(-\frac{\tau}{2}(X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)\bigg) (2\pi)^{-\frac{p_0}{2}} (\tau)^{\frac{p_0}{2}} a_0^{\frac{p_0}{2}} \\
&\qquad\qquad \times (\det((X_0'X_0)^{-1}))^{-\frac{1}{2}} I_A(a_0) (2\pi)^{-\frac{p_1}{2}} (\tau)^{\frac{p_1}{2}} \\
&\qquad\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)\bigg) \\
&\qquad\qquad \times (\det(\boldsymbol{\Lambda}_1^{-1}))^{-\frac{1}{2}} \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg) \\
&\qquad\qquad \times \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} (\tau)^{\alpha_0 - 1} \exp(-\gamma_0\tau)
\end{aligned}
$$

As this does not result in a recognizable distribution, full conditionals can be used to sample from the posterior. First, the following parameters are defined:

$$
\begin{aligned}
\boldsymbol{\Lambda}_{0c} &= (X_2'X_2) + a_0(X_0'X_0) \\
\boldsymbol{\mu}_{0c} &= \boldsymbol{\Lambda}_{0c}^{-1}\Big(X_2'Y + a_0X_0'Y_0 - X_2'X_1\boldsymbol{\beta}_1\Big) \\
\boldsymbol{\Lambda}_{1c} &= (X_1'X_1) + \boldsymbol{\Lambda}_1 \\
\boldsymbol{\mu}_{1c} &= \boldsymbol{\Lambda}_{1c}^{-1}\Big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big) \\
\alpha_c &= \frac{1}{2}\big(n + p\big) + \alpha_0 \\
\gamma_c &= \frac{1}{2}\Big[(X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y) + a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0) \\
&\qquad + (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\Big] + \gamma_0
\end{aligned}
$$

The full conditionals (see Appendix E for the derivations) are:

$$
\begin{aligned}
&\boldsymbol{\beta}_0 | \boldsymbol{\beta}_1, \tau, a_0, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{0c}, \tau^{-1}\boldsymbol{\Lambda}_{0c}^{-1}) \\
&\boldsymbol{\beta}_1 | \boldsymbol{\beta}_0, \tau, a_0, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{1c}, \tau^{-1}\boldsymbol{\Lambda}_{1c}^{-1}) \\
&\tau | \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D \sim \text{Gam}(\alpha_c, \gamma_c)
\end{aligned}
$$

$$a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D$$
$$\sim \text{TruncGam}\left(\frac{p_0}{2}+1,\ \ 2\tau^{-1}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)^{-1},\ \ 0,\ \ 1\right)$$

Unlike NPBPP, the full conditional for a_0 is a recognizable distribution (while a truncated gamma is not a commonly used distribution, it is still a known distribution that is easy to sample from; see Nadarajah and Kotz 2006). Therefore, a more computationally efficient Gibbs sampler can be used to sample from the posterior distribution, allowing for quicker interim analysis when incorporating DT information into OT.

4.2.2 Comparing Conditional Normalized Partial Borrowing Power Prior to Normalized Partial Borrowing Power Prior

While the purpose of full conditionals is to sample from the posterior distribution conditionally and iteratively, they are also a way to understand how model parameters are being updated. Therefore, the full conditionals when using CNPBPP and the full conditionals when using NPBPP can be compared to understand how the two priors influence the posterior.

Full Conditional		NPBPP	CNPBPP
$\boldsymbol{\beta}_0 \vert \boldsymbol{\beta}_1, \tau, a_0, D_0, D$	a	$\boldsymbol{\Lambda}_{0f}^{-1}\big(X_2'Y - X_2'X_1\boldsymbol{\beta}_1 + a_0X_0'Y_0\big)$	$\boldsymbol{\Lambda}_{0c}^{-1}\big(X_2'Y - X_2'X_1\boldsymbol{\beta}_1 + a_0X_0'Y_0\big)$
$\sim$ Normal(a, b)	b	$\tau^{-1}\boldsymbol{\Lambda}_{0f}^{-1} = \tau^{-1}\big((X_2'X_2) + a_0(X_0'X_0)\big)^{-1}$	$\tau^{-1}\boldsymbol{\Lambda}_{0c}^{-1} = \tau^{-1}\big((X_2'X_2) + a_0(X_0'X_0)\big)^{-1}$
$\boldsymbol{\beta}_1 \vert \boldsymbol{\beta}_0, \tau, a_0, D_0, D$	c	$\boldsymbol{\Lambda}_{1f}^{-1}\big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\big)$	$\boldsymbol{\Lambda}_{1c}^{-1}\big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\big)$
$\sim$ Normal(c, d)	d	$\tau^{-1}\boldsymbol{\Lambda}_{1f}^{-1} = \tau^{-1}\big((X_1'X_1) + \boldsymbol{\Lambda}_1\big)^{-1}$	$\tau^{-1}\boldsymbol{\Lambda}_{1c}^{-1} = \tau^{-1}\big((X_1'X_1) + \boldsymbol{\Lambda}_1\big)^{-1}$
	e	$\frac{n+n_0a_0+p_1}{2} + \alpha_0$	$\frac{1}{2}\big(n+p\big) + \alpha_0$
$\tau \vert \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D$	f	$\frac{1}{2}\big[(X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)$	$\frac{1}{2}\big[(X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)$
$\sim$ Gam(e, f)		$+a_0(X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0)$	$+a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)$
		$+(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\big] + \gamma_0$	$+(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\big] + \gamma_0$
$a_0 \vert \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D$	g	NA	$\frac{p_0}{2} + 1$
$\sim$ TG(g, h, i, j)	h	NA	$2\tau^{-1}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)^{-1}$
where $A = [i, j]$	A	For $\alpha_0 < \frac{p_0}{2}$, $\big[\frac{p_0 - 2\alpha_0}{n_0}, 1\big]$; else, $[0, 1]$	$[0, 1]$

Table 4.1: Full Conditionals When Using NPBPP and When Using CNPBPP

and the full conditional for a_0 using NPBPP is

$$p(a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D) \propto (a_0)^{\frac{p_0}{2}} \frac{\gamma_{0n}^{\alpha_{0n}}}{\Gamma(\alpha_{0n})} (\tau)^{(\frac{n_0a_0}{2} + \alpha_0) - 1} I_A(a_0) \exp\left(- \frac{a_0\tau}{2} \Big((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) \Big) \right)$$

As can be seen in Table 4.1, the full conditionals for $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are the same, conditioned on τ and a_0, regardless of whether NPBPP or CNPBPP was used. Therefore, the full conditionals for $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are informing the posterior in the same way, conditioned on τ and a_0 being the same.

The differences in full conditionals for τ are driven solely by whether τ was integrated out of the denominator of the NPP (as in NPBPP) or not (as in CNPBPP). Of note, consider the rate parameter for the two gamma distributions in Table 4.1 (referred to previously as γ_f and γ_c). R. Christensen (2020) show that

$$(X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0-Y_0) = (X_0\boldsymbol{\beta}_0-M_0Y_0)'(X_0\boldsymbol{\beta}_0-M_0Y_0)+(M_0Y_0-Y_0)'(M_0Y_0-Y_0),$$

where M_0 is the perpendicular projection operator on $C(X_0)$. Furthermore, $(M_0Y_0 - Y_0)'(M_0Y_0 - Y_0)$ is the sum of squares error for Y_0 (SSE_0). Therefore,

$$\begin{aligned} a_0(X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) & \\ &= a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0) + a_0(M_0Y_0 - Y_0)'(M_0Y_0 - Y_0) \\ &= a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0) + a_0(\text{SSE}_0) \end{aligned}$$

Therefore, using Table 4.1, $\gamma_c = \gamma_f - \frac{1}{2}(a_0 \times \text{SSE}_0)$; that is to say, the rate parameter for the full conditional on τ when using CNPBPP is the rate parameter for the full conditional on τ when using NPBPP minus one-half of a down-weighted sum of squares error for Y_0.

Finally, the full conditionals for a_0 are different by design. How the full conditionals for τ and a_0 affect the (small-sample) posterior can be seen visually when considering a specific example, and can be found in Section 4.3 for the electric semi-truck example.

The full conditionals show that the (small-sample) posterior distribution will be different when using NPBPP compared to CNPBPP. However, any reasonable prior should lead to the same posterior inference when enough data from OT has been

obtained (BIDA). Therefore, asymptotically and conditioned on a_0, the influence of these two priors on the posterior should decrease as the amount of OT data increases (BDA). This is an informal result (for constructs that satisfy the assumptions) of the Bernstein-von Mises theorem: "in large samples the effect of the prior density π disappears: 'the data overwhelms the prior' " (Johnstone 2010). Therefore, the influence of the prior on the marginal posteriors of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and τ is typically diminished as the sample size increases.

Now consider a_0. When $a_0 = 1$, DT data and OT data are fully pooled; as the amount of OT data increases, DT data will eventually be overwhelmed by OT data. Since a_0 is bounded between 0 and 1, this will hold for any value a_0 takes on, with enough OT data. When a_0 is fixed, it is likely that the two posteriors would, asymptotically, be the same. However, when a_0 is not fixed, no amount of OT data will ever fully overwhelm the prior on a_0. This can be seen when considering the full conditional for a_0—which does not directly get updated by the OT data, Y, but only by the DT data, Y_0. While the full conditional for a_0 will update as more OT data are obtained (due to being conditioned on β_0 and τ which are updated by the OT data), the prior will not be fully overwhelmed and will exert some influence on the posterior. Asymptotic equivalence and the properties of a_0 and τ when using NPBPP compared to using CNPBPP are areas of further exploration. However, even for small samples, CNPBPP remains a valid alternative to NPBPP.

4.3 Implementing the Process

This section considers the case where DT information is both available and incorporated into OT through CNPBPP. This section also compares CNPBPP to NPBPP when all operational test events have been completed. As was done for the OT data sets used in Chapters 2 and 3, simulated data sets, with various distributions of the

DT response, were generated to provide insight into how CNPBPP performs across a range of possible scenarios.

Extending the electric semi-truck example from Chapters 2 and 3, consider a (well-designed) developmental test in which data was collected on the number of miles traveled on one charge, and that the DT model parameters were a subset of the OT model parameters. The DT data that was deemed functionally representative for inclusion in the CNPBPP had the following set-up: an electric semi-truck with a refrigerated payload type and a heavy weight was used in good weather to accomplish a 2^2 full factorial test design with five replicates where the only two factors considered were terrain and temperature (defined the same as in OT). Therefore, let $\boldsymbol{\beta}_1 = (\omega_2, \omega_3, \theta_2, \gamma_2, \alpha\theta_{(22)}, \alpha\gamma_{(22)}, \beta\theta_{(22)}, \beta\gamma_{(22)}, \theta\gamma_{(22)})$ and let $\boldsymbol{\beta}_0 = (\eta, \alpha_2, \beta_2, \alpha\beta_{(22)})$. The DT data model is then:

$$y_{0ijp} = \eta + \alpha_i + \beta_i + \alpha\beta_{(ij)} + \epsilon_{0ijp}$$

While a DT set-up at the hardest levels of the system is unlikely to be encountered in practice, this construct was assumed for consistency with previous chapters, so that direct comparisons can be made and to ensure that the model parameters have the same interpretation (specifically, so that η had the same interpretation).

Having established the data model, the initial priors are selected. Following Section 4.2.1, the initial priors are:

- $\boldsymbol{\beta}_1|\tau \sim \mathcal{N}(\boldsymbol{\mu}_1, \tau^{-1}\boldsymbol{\Lambda}_1)$ where $\boldsymbol{\mu}_1 = (-25, -50, 100, 100, 0, 0, 0, 0, 0)$ and $\boldsymbol{\Lambda}_1 = \frac{1}{n}(X_1'X_1)$
- $p_0(\boldsymbol{\beta}_0|\tau) \propto 1$
- $\tau \sim \text{Gam}(0.0001, 0.0001)$
- $a_0 \sim \text{Beta}(1, 1)$

The prior on $\boldsymbol{\beta}_1|\tau$ is a unit information prior developed by Kass and Wasserman (1995)—a special case of the g prior with a strength of one prior observation centered at $\boldsymbol{\mu}_1$ (Liang et al. 2008). Further, $\boldsymbol{\mu}_1$ was selected to be the mean of the weakly informative priors selected in Chapters 2 and 3.

For this chapter, a subset of the 21 OT data sets from Chapters 2 and 3 are used—specifically data sets 1, 2, 3, 10, and 11. These data sets were selected based on the results from Chapter 3. Recall that OT data set 1 ended OT early based on PP for all n_{obs}, evaluating the measure as not met. OT data set 2 was selected for similar reasons; however, if PP was only calculated when $n_{obs} = 60$ and 75 then OT would not have ended early based on PP. OT data set 3 ended OT early based on PP, evaluating the measure as met. Both OT data set 10 and 11 were unable to end test early based on PP, requiring all test events to be accomplished. However, using OT data set 10 resulted in evaluating the measure as not met using posterior probability while using OT data set 11 resulted in evaluating the measure as met. These data sets provide an avenue for examining how edge cases are affected by different DT data sets when testing can end early using PP (either for evaluating the measure as met or evaluating it as not met) and when testing requires all test events to be seen (either ultimately evaluating the measures as met or evaluating it as not met).

The following provides a description of the DT data sets used in this chapter. DT data set 1 was created in the same manner as the corresponding OT data set. For example, OT data set 2 was created using Appendix B with $\eta = 345$ and OT data set 3 was created with $\eta = 347$; therefore, DT data set 1 for OT data set 2 was also created using $\eta = 345$ while DT data set 1 for OT data set 3 was created using $\eta = 347$. DT data set 2 was created in the same manner as the corresponding OT data set, but increased η by 5. Therefore, DT data set 2 for OT data set 2 was created using $\eta = 350$. DT data set 3 was created in the same manner as the

corresponding OT data set, but increased η by 150. DT data set 4 was created to examine how changing the true parameter values for α_2 and β_2 (while maintaining the same sum of the two parameters) would affect the analysis. In all OT data sets, $\alpha_2 = 50$ and $\beta_2 = 15$; in every DT data set 4, $\alpha_2 = 75$ and $\beta_2 = -10$. Therefore, when both α_2 and β_2 are observed (or, when neither are observed) the expectation should be the same for OT and DT; furthermore, in a balanced design (such as is presented here), the expectation of the average miles traveled should be the same. Finally, DT data set 5 was created in the same manner as the corresponding OT data set (as in DT data set 1); however, a smaller standard deviation was used for DT data set 5 compared to the corresponding OT data set. Using notation from Appendix B, DT data set 5 was generated by $b = 1$ and $s = 0$ (when paired with OT data sets 1, 2, and 3) or $b = 50$ and $s = 0$ (when paired with OT data sets 10 and 11), compared to the OT data sets that was generated by $b = 50$ / $s = 2$ and $b = 100$ / $s = 9$ respectively. DT data set 1 was selected to explore how much information would be borrowed from DT in the (unlikely) case where DT and OT data are exactly the same. DT data set 4 was selected to provide insights into how sensitive CNPBPP is to a change in the sign of a parameter when the expectation remained largely the same compared to OT. Finally, DT data sets 2, 3, and 5 were all selected to explore how robust CNPBPP is to deviations in assumptions that could occur in practice. Given that DT is more controlled than OT, DT can see better system performance than OT—DT data set 2 is the case where DT data is slightly better than in OT; DT data set 3 is the case where DT data is much better than in OT; and DT data set 5 is the case where the variance is artificially reduced in DT compared to OT.

The results for each data set combination can be found in Tables 4.2 and 4.3, as well as in Figures 4.2–4.6. PP was calculated in R using the same two-stage sampling algorithm detailed in Chapter 3, and was also based on $n_j = 18,000$ nested samples and $n_i = 1,000$ outer samples (after examining both ACF and trace plots,

and removing burn-in samples, as well as for comparison purposes with Chapters 2 and 3). Recall from Chapter 3 that n_f is the number of observations required to be seen before calculating PP. Chapter 3 discussed that it was possible to make the wrong decision due to a lack of information when $n_f = 0$ if reference priors are used; as such Chapter 3 considered $n_{obs} = 45, 60, 75$, as is done in Table 4.2 for comparison purposes. When using CNPBPP, selecting $n_f > 0$ ensures that OT data is obtained, and the analysis is not completely driven by DT data. Finally, recall from Chapter 3 that $\theta_U = 0.95$ and $\theta_L = 0.05$.

As can be seen in Tables 4.2 and 4.3, for any result at a given n_{obs}, the posterior expectation of a_0 is consistent within an OT data set and DT data set paring. This demonstrates that conducting interim analysis, when all observations have not been seen, does not influence the posterior expectation of a_0. This implies that a_0 behaves consistently within our proposed method.

Table 4.2 contains results for OT data sets 1, 2, and 3, which were generated using the smallest error transformation. In Chapter 3, OT data set 1 would have allowed for testing to end after $n_{obs} = 45$, evaluating the measure as not met using PP; Table 4.2a demonstrates that OT data set 1 would still allow for testing to be ended early and evaluate the measure as not met using PP at $n_{obs} = 45$ for every DT data set, with the exception of DT data set 3. The PP for DT data set 3 paired with OT data set 1 is much higher than the other pairings, requiring all test events to be seen before evaluating the measure. This illustrates that when prior data is incorporated into analysis that is dissimilar, more testing may be required compared to using weakly informative priors. In Table 4.2b, using DT data sets 1 or 4 results in the same conclusion about the measure as using weakly informative priors in Chapters 2 and 3, demonstrating the same borderline conclusion about the measure. However, DT data sets 2 and 3 are examples of better performance in DT than OT, influencing conclusions. While DT data set 2 has more influence on the

posterior analysis, DT data set 3 has a much better system performance than OT—which in-turn has a greater influence on the PP, in spite of a small a_0. Considering Table 4.2c, all operational tests are allowed to stop much earlier when using DT information, compared to when using weakly informative priors.

Next, consider Table 4.3, where OT data sets 10 and 11 have a larger error transformation compared to Table 4.2. In Chapter 3, OT data set 10 required seeing all test events before testers could make a conclusion about the measure (not met) when using a weakly informative prior; in Table 4.3a, using any of the DT data sets as a prior would allow testers to end an operational test early, with similar outcomes as in Table 4.2a. Similarly, in Chapter 3, OT data set 11 required seeing all test events before testers could make a conclusion about the measure when using a weakly informative prior; while evaluated as met, the posterior probability was very close to θ_T. In Table 4.3b, using DT data sets 1 and 3, testers can end OT early based on PP; alternatively, DT data sets 2, 4, and 5 do not allow for ending OT early based on PP. This implies that our proposed method for using PP (and using CNPBPP) will result in ending OT early when the posterior probability is not very close to θ_T.

Additionally, consider a_0 in Tables 4.2 and 4.3 for DT data set 3. Note that in Table 4.2, when the smallest error transformation is used to generate the data sets, a_0 is approximately 0.04 for any n_{obs} or OT data set. In Table 4.3, when a larger error transformation is used to generate the data sets, a_0 increases to approximately 0.3. While not shown, when using the data sets with the largest error transformation, a_0 continues to increase (to approximately 0.5). As the OT data becomes more diffuse, DT data set 3 becomes more commensurate with (less disinformative relative to) the OT data set it is paired with, which in-turn increases a_0. However, we do not see this increase in a_0 for any other DT data sets, which is due to increasing the errors surrounding the fitted values without also scaling the fitted values when generating the data sets.

Prior		WIP*	CNPBPP Using DT Data Set...									
			1		2		3		4		5	
MNP*	n_{obs}	PP	PP	a_0	PP	a_0	PP	a_0	PP	a_0	PP	a_0
8	45	$>0.0001^{\dagger}$	$>0.0001^{\dagger}$	0.6528	0.0010	0.6504	0.6076	0.0426	$>0.0001^{\dagger}$	0.6948	0.0150	0.6961
13	60	$>0.0001^{\dagger}$	$>0.0001^{\dagger}$	0.6347	$>0.0001^{\dagger}$	0.6563	0.5796	0.0440	$>0.0001^{\dagger}$	0.6980	0.0080	0.7026
18	75	$>0.0001^{\dagger}$	$>0.0001^{\dagger}$	0.6501	0.0020	0.6601	0.5986	0.0424	$>0.0001^{\dagger}$	0.6932	0.0210	0.6932
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		0.7856	0.7847	0.6419	0.7894	0.6537	0.8000	0.0425	0.7854	0.6922	0.7932	0.6950

(a) OT Data Set 1 ($\eta = 343$)

Prior		WIP*	CNPBPP Using DT Data Set...									
			1		2		3		4		5	
MNP*	n_{obs}	PP	PP	a_0	PP	a_0	PP	a_0	PP	a_0	PP	a_0
8	45	0.0412	0.2262	0.6338	0.6186	0.6541	$>0.9999^{\dagger}$	0.0419	0.2072	0.6920	0.2032	0.6943
13	60	0.1116	0.2292	0.6395	0.5976	0.6624	$>0.9999^{\dagger}$	0.0428	0.2182	0.6921	0.1972	0.6925
18	75	0.1781	0.2272	0.6514	0.6306	0.6553	$>0.9999^{\dagger}$	0.0426	0.2132	0.7062	0.2012	0.6906
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		0.7968	0.7968	0.6418	0.8005	0.6542	0.8088	0.0424	0.7959	0.6930	0.8040	0.6938

(b) OT Data Set 2 ($\eta = 345$)

Prior		WIP*	CNPBPP Using DT Data Set...									
			1		2		3		4		5	
MNP*	n_{obs}	PP	PP	a_0	PP	a_0	PP	a_0	PP	a_0	PP	a_0
8	45	0.6281	0.9970	0.6425	$>0.9999^{\dagger}$	0.6581	$>0.9999^{\dagger}$	0.0424	0.9960	0.7014	$>0.9999^{\dagger}$	0.6974
13	60	0.8702	0.9920	0.6599	0.9990	0.6368	$>0.9999^{\dagger}$	0.0420	0.9950	0.6974	$>0.9999^{\dagger}$	0.6938
18	75	0.9748	0.9950	0.6398	$>0.9999^{\dagger}$	0.6511	$>0.9999^{\dagger}$	0.0433	0.9970	0.6905	$>0.9999^{\dagger}$	0.6883
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		0.8059	0.8078	0.6400	0.8088	0.6558	0.8180	0.0424	0.8052	0.6921	0.8123	0.6953

(c) OT Data Set 3 ($\eta = 347$)

* WIP is weakly informative prior (Ch 3 results); MNP is min. number of obs. informing posterior of each parameter.
† Computationally, this was numerically indistinguishable from 0 or 1, but not actually 0 or 1.

Table 4.2: PP and $\Pr_{\phi|X}(\phi \geq 400)$ for Various n_{obs}s and ηs

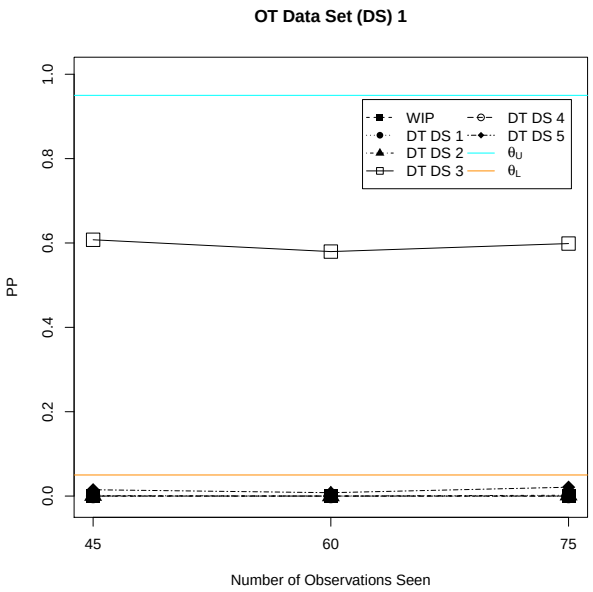

Figure 4.2: PP for OT Data Set 1

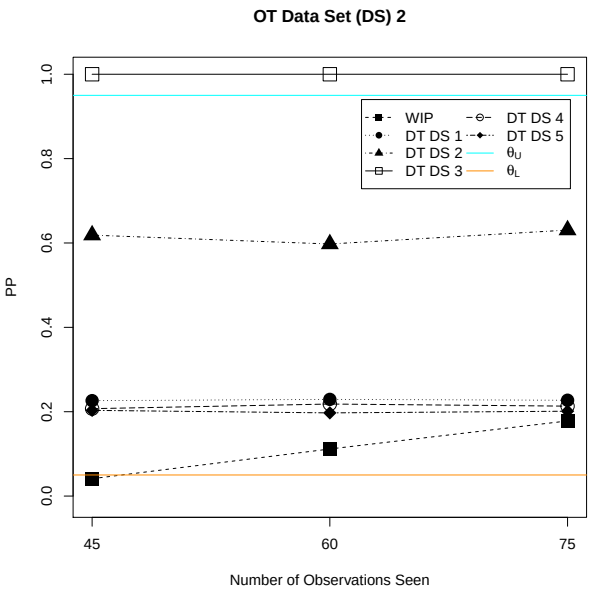

Figure 4.3: PP for OT Data Set 2

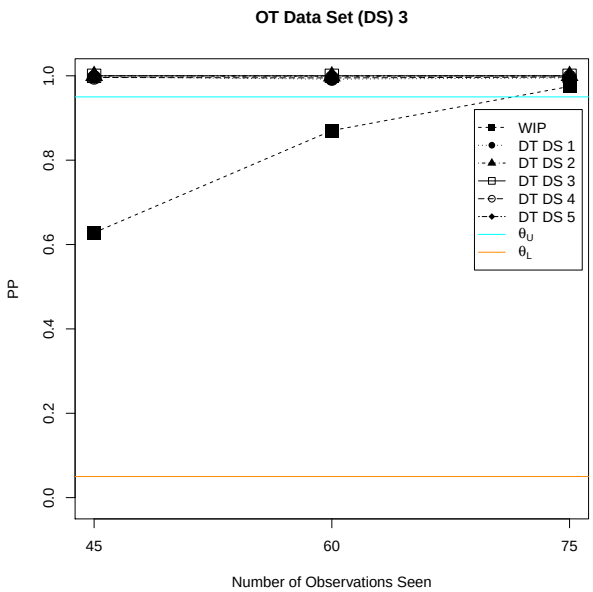

Figure 4.4: PP for OT Data Set 3

Prior		WIP*	CNPBPP Using DT Data Set...									
			1		2		3		4		5	
MNP*	n_{obs}	PP	PP	a_0	PP	a_0	PP	a_0	PP	a_0	PP	a_0
8	45	0.0633	$<0.0001^{\dagger}$	0.6639	$<0.0001^{\dagger}$	0.6499	$>0.9999^{\dagger}$	0.2972	0.0010	0.6648	0.0010	0.6789
13	60	0.2968	0.0010	0.6351	0.0010	0.6528	$>0.9999^{\dagger}$	0.3005	0.0010	0.6729	$<0.0001^{\dagger}$	0.6681
18	75	0.1861	$<0.0001^{\dagger}$	0.6550	$<0.0001^{\dagger}$	0.6582	$>0.9999^{\dagger}$	0.2974	$<0.0001^{\dagger}$	0.6708	0.0020	0.6771
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		0.7959	0.7857	0.6475	0.7898	0.6514	0.8375	0.2917	0.7864	0.6612	0.7901	0.6761

(a) OT Data Set 10 ($\eta = 347$)

Prior		WIP*	CNPBPP Using DT Data Set...									
			1		2		3		4		5	
MNP*	n_{obs}	PP	PP	a_0	PP	a_0	PP	a_0	PP	a_0	PP	a_0
8	45	0.5634	0.0350	0.6618	0.1802	0.6430	$>0.9999^{\dagger}$	0.2913	0.1161	0.6555	0.2633	0.6807
13	60	0.7918	0.0440	0.6414	0.2152	0.6524	$>0.9999^{\dagger}$	0.2946	0.1111	0.6588	0.2843	0.6738
18	75	0.9014	0.0360	0.6389	0.2392	0.647	$>0.9999^{\dagger}$	0.2964	0.1161	0.6549	0.2683	0.6944
$\Pr_{\phi\|X}(\phi \geq \phi_0)$		0.8018	0.7932	0.6473	0.7943	0.6509	0.8443	0.2951	0.7952	0.6601	0.7970	0.6788

(b) OT Data Set 11 ($\eta = 349$)

* WIP is weakly informative prior (Ch 3 results); MNP is min. number of obs. informing posterior of each parameter.
† Computationally, this was numerically indistinguishable from 0 or 1, but not actually 0 or 1.

Table 4.3: PP and $\Pr_{\phi|X}(\phi \geq 400)$ for Various n_{obs}s and ηs

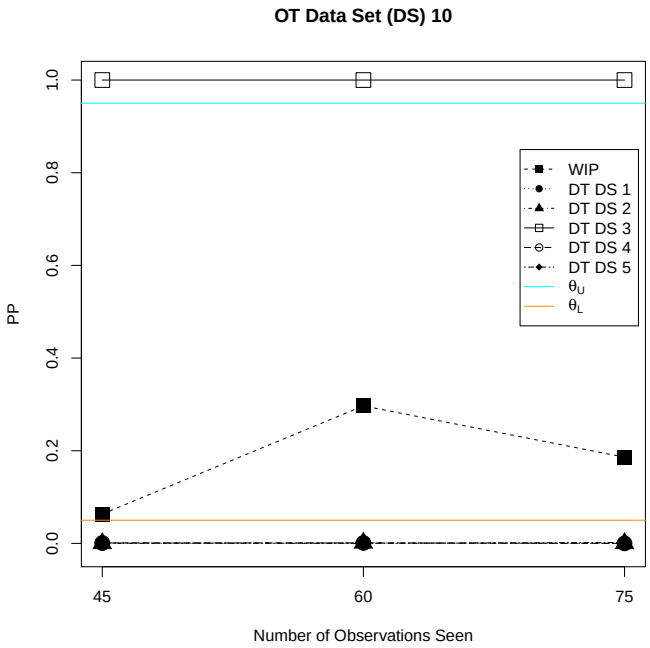

Figure 4.5: PP for OT Data Set 10

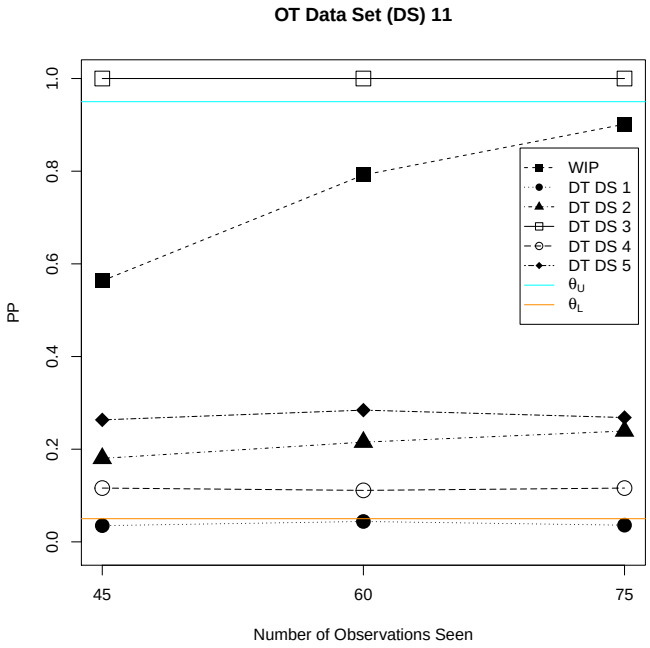

Figure 4.6: PP for OT Data Set 11

As discussed in Chapter 2, model parameter estimation is not a focus of the methods presented here. However, the standard deviation of those model parameter estimates is an important aspect of being able to make decisions about a measure. Recall from Chapter 3 that the sampling method for PP requires a two stage sampling method, where the nested loop ultimately calculates $\Pr(\phi \geq \phi_0|X, Y) \geq \theta_T$ and the outer loop samples from the predictive distribution of Y. When the standard deviation of model parameter estimates decrease, it corresponds to seeing less varied posterior draws for the model parameters. This decreases the variation seen in the Ys drawn in the outer loop of the sampling method, resulting in less varied data being seen in the nested loop. The conditional draws in the nested loop then more accurately estimate the $(1 - \theta_T)$ quantile, which in term produces a more accurate PP. As PP less becomes influenced by the variability in the nested loop's conditional draws, quicker decisions (i.e., decisions made about a measure based on fewer observational units) can be made. This can be seen in Table 4.4b for OT data set 3. When a_0 allows for more borrowing of DT information (i.e. DT data sets 1, 2, 4, and 5), the posterior standard deviation for model parameters decreases. Alternatively, when a_0 does not allow for more borrowing of DT, as in DT data set 3, the posterior standard deviation is comparable to using weakly informative priors. Therefore, compared to weakly informative priors, using CNPBPP allows for the possibility that the posterior standard deviation can be smaller when the data is commensurate, while not increasing the standard deviation when the data is not commensurate. This ultimately allows PP to be more accurate, which, and as Table 4.2c demonstrates, leads to making quicker decisions.

Model Parameter	True Value	WIP*	CNPBPP Using DT Data Set...				
			1	2	3	4	5
η	347	352.75	349.54	350.30	363.89	349.61	350.99
α_2	50	60.07	53.15	54.03	58.55	62.02	58.28
β_2	15	15.49	11.04	14.03	9.44	6.51	12.32
ω_2	0	-6.62	-0.47	-2.11	-8.82	0.34	-2.28
ω_3	-5	-4.01	2.12	1.05	-4.81	1.68	0.43
γ_2	75	66.27	66.43	65.70	54.91	64.75	64.47
δ_2	50	56.92	59.65	58.60	47.15	58.25	57.48
$(\alpha\beta)_{(22)}$	20	-12.61	-1.32	-2.71	-10.92	-12.25	-6.89
$(\alpha\gamma)_{(22)}$	50	38.54	41.27	40.85	39.43	39.17	39.25
$(\alpha\delta)_{(22)}$	25	17.15	16.54	16.90	18.07	14.13	15.48
$(\beta\gamma)_{(22)}$	50	65.56	67.04	65.30	70.46	73.64	67.83
$(\beta\delta)_{(22)}$	25	15.80	13.89	12.41	18.62	20.48	14.91
$(\gamma\delta)_{(22)}$	25	22.22	19.12	21.50	32.22	18.09	22.26
$\tau^{\dagger}$	—	0.414	0.470	0.473	0.446	0.480	0.480
a_0	NA	NA	0.640	0.656	0.042	0.692	0.695

(a) Posterior Expectation for Model Parameters

Model Parameter	WIP*	CNPBPP Using DT Data Set...				
		1	2	3	4	5
η	18.82	14.97	14.83	19.69	14.66	14.59
α_2	21.30	17.90	17.75	21.07	17.34	17.52
β_2	19.73	17.80	17.64	21.13	17.47	17.35
ω_2	15.27	13.93	13.79	15.61	13.72	13.66
ω_3	12.93	11.91	11.77	13.02	11.76	11.69
γ_2	21.08	19.10	18.97	21.27	18.88	18.85
δ_2	21.66	19.33	19.27	21.83	19.08	18.99
$(\alpha\beta)_{(22)}$	21.16	19.77	19.58	21.48	19.23	19.15
$(\alpha\gamma)_{(22)}$	21.69	19.54	19.50	21.04	19.28	19.41
$(\alpha\delta)_{(22)}$	22.22	20.09	19.95	21.66	19.76	19.83
$(\beta\gamma)_{(22)}$	21.23	19.71	19.55	21.16	19.32	19.39
$(\beta\delta)_{(22)}$	21.32	19.73	19.63	21.25	19.52	19.43
$(\gamma\delta)_{(22)}$	23.22	20.75	20.68	23.30	20.57	20.42
$\tau^{\dagger}$	0.0707	0.0747	0.075	0.0738	0.0758	0.076
a_0	NA	0.2323	0.2288	0.0289	0.2178	0.2162

(b) Posterior Standard Deviation for Model Parameters

* WIP is the Weakly Informative Prior used in Chapters 2 and 3

† For ease of display, results for τ are multiplied by $1,000$; see Appendix B for the transformation on τ to obtain the true value.

Table 4.4: Posterior Expectations and Standard Deviations for Model Parameter Based on $n = 80$ Observations for OT Data Set 3

As discussed in Section 4.1.4, Neelon and O'Malley (2010) suggest that using a NPP (which is incorporated into CNPBPP) would need a fairly informative prior on a_0 to bound a_0 away from 0, due to the tendency of NPP to overly down-weight historical information. While the research area of OT is unlikely to find issue with a prior that overly down-weights DT data, other areas of research may be interested in the use of a more informative prior on a_0. Previously, it was stated that the Beta(1,1) prior on a_0 used thus far was selected to ensure conjugacy; alternatively, a choice of Beta$(d, 1)$ for any $d > 0$ could have been selected while retaining recognizable full conditionals. Tables 4.5 and 4.6 demonstrate how a_0 changes with the choice of d—specifically, how the posterior expectation of a_0 increases as its prior becomes more informative. Therefore, as the prior on a_0 becomes more informative, CNPBPP is able to incorporate more DT data. As demonstrated in Table 4.5, this will further decrease the posterior standard deviation of the model parameters, when the data is commensurate. However, if the data is not commensurate, as is shown in Table 4.6, posterior standard deviations can become worse. Furthermore, if the prior on a_0 is too strong and there is limited OT data, the differences between the two data sets cannot overwhelm the prior on a_0. As can be seen in Table 4.6, this can create a disinformative prior, and lead to misplaced confidence in a system obtaining the required threshold. Therefore, even though a Beta$(1, 1)$ prior on a_0 my overly down-weight historical information, we contend that a Beta$(1, 1)$ prior is better for use in OT than a stronger prior that may become disinformative.

Model	True	Beta(0.2, 1)		Beta(1,1)		Beta(5,1)		Beta(50,1)	
Parameter	Value	E(·)*	sd(·)*	E(·)*	sd(·)*	E(·)*	sd(·)*	E(·)*	sd(·)*
η	345	348.10	15.48	347.77	14.93	346.55	14.02	346.04	13.60
α_2	50	53.89	18.33	52.80	17.90	51.28	17.11	50.35	16.70
β_2	15	10.82	18.22	10.94	17.78	11.45	17.04	11.63	16.66
ω_2	0	−0.72	14.07	−0.46	13.89	0.15	13.72	0.44	13.71
ω_3	−5	1.78	11.88	2.04	11.88	2.76	11.81	3.13	11.78
γ_2	75	65.89	19.23	66.40	19.23	67.59	18.77	68.20	18.79
δ_2	50	59.11	19.48	59.56	19.21	60.87	18.96	61.40	18.79
$(\alpha\beta)_{(22)}$	20	−2.47	19.89	−1.09	19.74	1.22	19.17	2.75	18.93
$(\alpha\gamma)_{(22)}$	50	41.14	19.64	41.30	19.70	41.74	19.42	41.84	19.27
$(\alpha\delta)_{(22)}$	25	16.54	20.17	16.89	20.07	16.79	19.74	16.95	19.66
$(\beta\gamma)_{(22)}$	50	67.50	19.77	67.02	19.72	65.95	19.43	65.40	19.31
$(\beta\delta)_{(22)}$	25	14.37	19.93	13.84	19.77	12.66	19.57	11.91	19.36
$(\gamma\delta)_{(22)}$	25	19.57	20.90	19.05	20.73	18.11	20.60	17.72	20.40
$\tau^{\dagger}$	—	0.473	0.0745	0.471	0.0744	0.467	0.0738	0.469	0.0739
a_0	NA	0.5408	0.2567	0.6418	0.2321	0.8447	0.1306	0.9804	0.0192
$\Pr_{\phi\|X}(\phi > \phi_0)$		0.7971		0.7968		0.7961		0.7948	

* E(·) is the posterior expectation of a model parameter, and sd(·) is the posterior standard deviation of a model parameter.
† For ease of display, results for τ are multiplied by $1,000$; see Appendix B for the transformation on τ to obtain the true value.

Table 4.5: Changing the Prior on a_0 for OT Data Set 2 and DT Data Set 1

Model	True	Beta(0.2, 1)		Beta(1,1)		Beta(5,1)		Beta(50,1)	
Parameter	Value	E(·)*	sd(·)*	E(·)*	sd(·)*	E(·)*	sd(·)*	E(·)*	sd(·)*
η	345	359.73	19.370	362.17	19.94	379.95	23.19	434.76	18.42
α_2	50	58.50	21.02	58.26	21.19	56.69	22.27	52.90	22.67
β_2	15	9.15	20.96	9.21	21.06	11.54	22.19	18.49	22.53
ω_2	0	−7.45	15.30	−8.96	15.62	−20.07	18.02	−54.92	18.48
ω_3	−5	−3.75	12.76	−4.86	13.06	−13.14	14.81	−38.85	15.87
γ_2	75	56.42	21.07	54.77	21.57	42.00	24.24	2.91	25.24
δ_2	50	48.93	21.47	46.99	21.96	32.79	24.90	−10.83	25.56
$(\alpha\beta)_{(22)}$	20	−10.78	21.26	−10.85	21.46	−12.61	22.99	−16.73	25.57
$(\alpha\gamma)_{(22)}$	50	39.61	20.91	39.56	21.14	39.27	22.73	37.08	26.29
$(\alpha\delta)_{(22)}$	25	17.81	21.64	18.24	21.93	22.37	23.59	34.11	26.72
$(\beta\gamma)_{(22)}$	50	70.75	21.07	70.58	21.15	68.56	22.75	62.07	26.19
$(\beta\delta)_{(22)}$	25	18.66	21.14	18.84	21.18	18.56	22.87	17.39	26.25
$(\gamma\delta)_{(22)}$	25	29.92	22.93	32.18	23.47	49.40	26.95	102.81	27.65
$\tau^{\dagger}$	—	0.457	0.0743	0.446	0.0739	0.379	0.0709	0.253	0.0401
a_0	NA	0.0291	0.0222	0.0424	0.0289	0.1674	0.1138	0.9773	0.0222
$\Pr_{\phi\|X}(\phi > \phi_0)$		0.8076		0.8088		0.8287		0.8819	

* E(·) is the posterior expectation of a model parameter, and sd(·) is the posterior standard deviation of a model parameter.
† For ease of display, results for τ are multiplied by $1,000$; see Appendix B for the transformation on τ to obtain the true value.

Table 4.6: Changing the Prior on a_0 for OT Data Set 2 and DT Data Set 3

Finally, Table 4.7 and Figures 4.7–4.10 compares the difference in posterior probabilities and a_0 when using CNPBPP and using NPBPP. For any OT data set, DT data sets 1, 2, 4, and 5 have posterior probabilities for CNPPBP that are within 0.002 of the posterior probabilities for NPBPP (in no consistent direction). Of the 25 data combinations considered in this table, only one data set combination results in an inconsistent evaluation between the two methods (OT data set 2 with DT data set 2). This demonstrates that it is likely that the more computationally efficient CNPBPP will largely lead to the same conclusions regarding measure evaluation as NPBPP, unless the posterior probability is very close to θ_T.

In all cases, DT data set 5 for all OT data sets demonstrates that NPBPP is more sensitive to deviations in τ than CNPBPP; however, that sensitivity to deviations in τ decreases as the the variance in the errors increases (i.e. a_0 is approximately 0.23 under a small variance using error transformation 1, but 0.47 under a larger variance using error transformation 2). As a_0 is not as driven by τ in CNPBPP, a_0 is not as sensitive to deviations in τ compared to the a_0 when using NPBPP. Therefore, given the OT and DT data sets were generated from the same true model parameters (except τ), CNPBPP is more robust to deviations in modeling assumptions about τ than NPBPP.

Furthermore, by considering DT data set 3 for all OT data sets, Table 4.7a demonstrates that NPBPP can lead to a more disinformative prior because the Gamma$(0.0001, 0.0001)$ prior on τ restricts a_0 so that $a_0 \in [0.2, 1]$—this is in contrast to CNPBPP, which allows for $a_0 \in [0, 1]$. For this example, the shape must be greater than or equal to 2 before $a_0 \in [0, 1]$ for NPBPP. Alternatively, as can be seen in Table 4.7b, as the variance in the errors increases, a_0 becomes less sensitive to the deviations in the DT data set when using NPBPP compared to using CNPBPP. This indicates that CNPBPP may not only be more robust in cases where τ differs in DT and OT, but also when the true model parameters differ in DT and OT.

OT Data Set	DT Data Set	CNPBPP $\Pr_{\phi\|X}(\phi > \phi_0)$	CNPBPP a_0	NPBPP $\Pr_{\phi\|X}(\phi > \phi_0)$	NPBPP a_0
1	1	0.7847	0.6419	0.7861	0.7298
	2	0.7894	0.6537	0.7896	0.7393
	3	0.8000	0.0425	0.8338	0.2848
	4	0.7854	0.6922	0.7841	0.7603
	5	0.7933	0.6950	0.7919	0.2312
2	1	0.7968	0.6418	0.7956	0.7272
	2	0.8005	0.6542	0.7994	0.7398
	3	0.8088	0.0424	0.8432	0.2832
	4	0.7959	0.6930	0.7955	0.7564
	5	0.8040	0.6938	0.8030	0.2314
3	1	0.8078	0.6400	0.8059	0.7285
	2	0.8088	0.6558	0.8085	0.7381
	3	0.8180	0.0424	0.8523	0.2844
	4	0.8053	0.6921	0.8070	0.7585
	5	0.8123	0.6953	0.8129	0.2312

(a) Error Transformation 1 (Small Variance)

OT Data Set	DT Data Set	CNPBPP $\Pr_{\phi\|X}(\phi > \phi_0)$	CNPBPP a_0	NPBPP $\Pr_{\phi\|X}(\phi > \phi_0)$	NPBPP a_0
10	1	0.7857	0.6475	0.7834	0.7256
	2	0.7898	0.6514	0.7878	0.7284
	3	0.8375	0.2917	0.8600	0.5298
	4	0.7864	0.6612	0.7887	0.7354
	5	0.7901	0.6761	0.7902	0.4722
11	1	0.7932	0.6473	0.793	0.7226
	2	0.7943	0.6509	0.7954	0.7284
	3	0.8443	0.2951	0.8667	0.5331
	4	0.7952	0.6601	0.7968	0.7356
	5	0.7970	0.6788	0.7994	0.4703

(b) Error Transformation 2 (Larger Variance than in Error Transformation 1)
* Definitions for Error Transformations 1 and 2 can be found in Appendix B.

Table 4.7: Comparison of Posterior Probability Results Using CNPBPP versus NPBPP

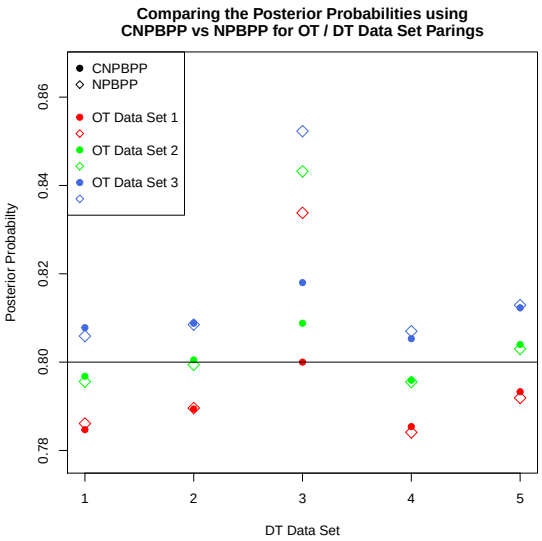

Figure 4.7: Posterior Probabilities

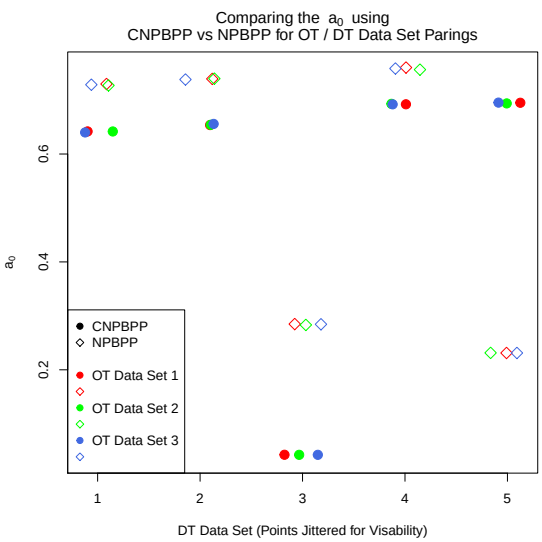

Figure 4.8: a_0

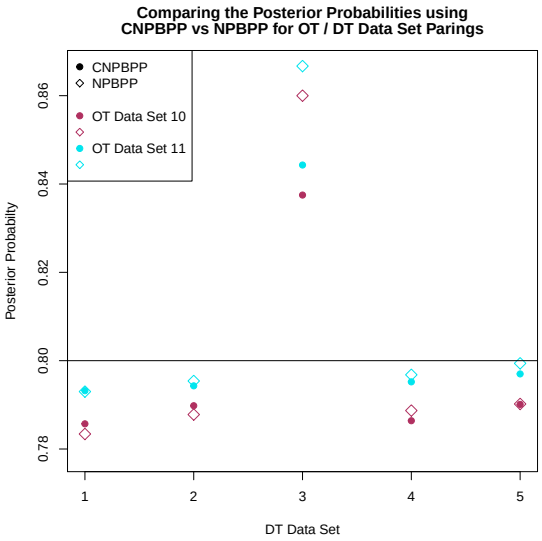

Figure 4.9: Posteiror Probabilities

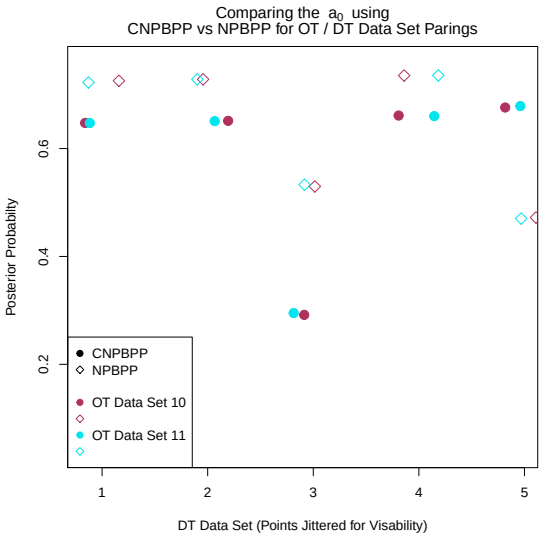

Figure 4.10: a_0

Figures 4.11 and 4.12 show the marginal posterior distribution for a_0 and τ when using a Gamma$(0.0001, 0.0001)$ prior on τ when DT data and OT data are the same (i.e. considering DT data set 1). With this paring, CNPBPP is less likely to borrow from DT than NPBPP, but the marginal posterior for τ are relatively similar.

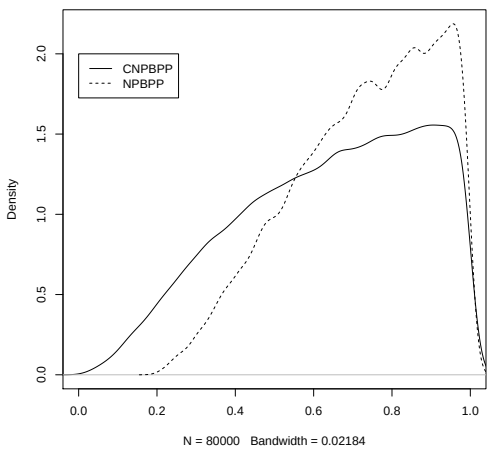

Figure 4.11: Marginal Posterior for a_0 comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 1 when the Initial Prior on τ was a Gamma$(0.0001, 0.0001)$

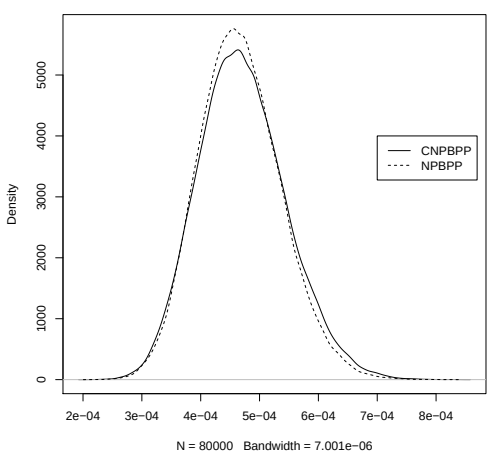

Figure 4.12: Marginal Posterior for τ comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 1 when the Initial Prior on τ was a Gamma$(0.0001, 0.0001)$

Now consider when DT data is much better than OT data (i.e. considering DT data set 3). Figures 4.13 and 4.14 show the marginal posterior distribution for a_0 and τ when using a Gamma$(0.0001, 0.0001)$ prior on τ. With this paring, CNPBPP is still less likely to borrow from DT than NPBPP, but the restriction NPBPP puts on a_0 is clearly seen. Furthermore, the marginal posteriors for τ are no longer similar.

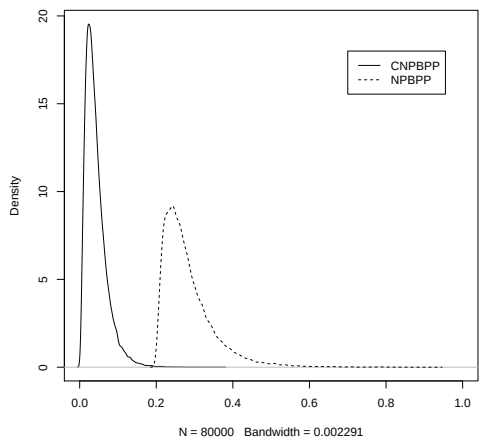

Figure 4.13: Marginal Posterior for a_0 comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 3 when the Initial Prior on τ was a Gamma$(0.001, 0.0001)$

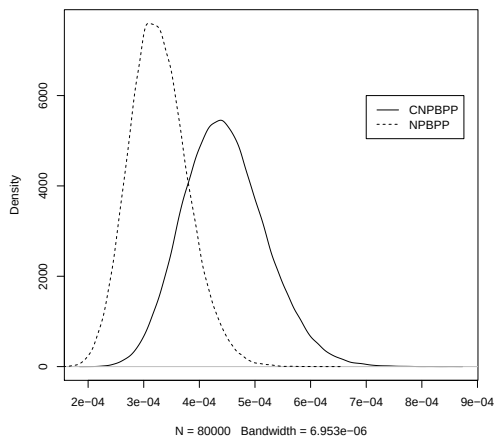

Figure 4.14: Marginal Posterior for τ comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 3 when the Initial Prior on τ was a Gamma$(0.0001, 0.0001)$

Alternately, consider when DT data and OT data are not the same, but the prior on τ is a Gamma$(2, 0.0001)$ which allows a_0 to have the same support when using CNPBPP or NPBPP ($a_0 \in [0, 1]$). Figures 4.15 and 4.16 show the marginal posterior distribution for a_0 and τ. With this prior, CNPBPP is still less likely to borrow from DT than NPBPP, but is now comparable to NPBPP. Furthermore, the marginal posteriors for τ are more similar than when using $\tau \sim$ Gamma$(0.0001, 0.0001)$.

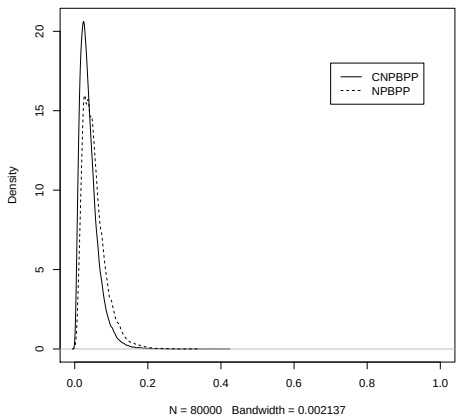

Figure 4.15: Marginal Posterior for a_0 comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 3 when the Initial Prior on τ was a Gamma$(2, 0.0001)$

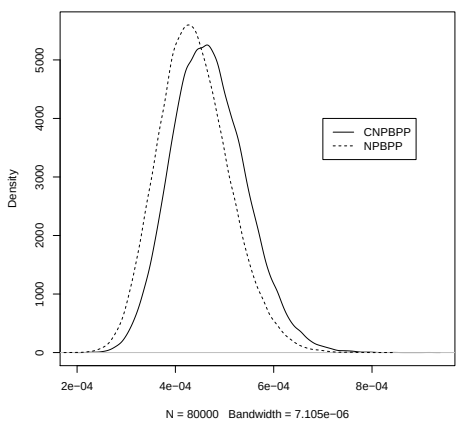

Figure 4.16: Marginal Posterior for a_0 comparing NPBPP vs CNPBPP for OT Data Set 2 and DT Data Set 3 when the Initial Prior on τ was a Gamma$(2, 0.0001)$

Ultimately, the decision to use CNPBPP or NPBPP relates to foundational issues, whether computational efficiency is required, and how robust the method needs to be. For the purposes of this research, using an NPBPP would require selecting prior on τ that ensures $a_0 \in [0, 1]$. If such a selection were not made, NPBPP can lead to a disinformative prior, and adversely affect measure evaluation. However, that selection must be counter-balanced against the implication of the selected prior on τ—as p_0 increases, a more informative prior on τ is required to ensure $a_0 \in [0, 1]$. Furthermore, NPBPP will not result in full conditionals that are fully recognizable, resulting in a more computationally inefficient sampling method. Alternatively, CNPBPP uses conditional probability to ensure that full conditionals are recognizable; therefore, not only does CNPBPP allow for $a_0 \in [0, 1]$ for normally distributed data regardless of the initial prior on τ, but it also ensures computational efficiency. Given that CNPBPP has protection built in against being a disinformative prior without creating a more informative prior on τ, is more computationally efficient, is robust to assumptions about τ, and does not result in substantial differences in measure evaluation, we believe CNPBPP is a more appropriate prior for OT.

Chapter 5

Discussion and Future Work

This research has demonstrated how the proposed methods can improve the effectiveness and efficiency of operational testing by stopping test early with confidence and incorporating developmental testing (DT) information into operational testing (OT) through the use of informative priors, providing both cost and schedule savings. Chapter 2 transitioned the OT paradigm into a Bayesian framework, and introduced mission sets; in doing so, Chapter 2 provided a means for using Bayesian methods within OT. Chapter 2 demonstrated that using the marginalized mission space provided more information about the variability in the parameter of interest among the operational environments in which the system will be used, finding that fewer systems should obtain ϕ_0 than would be considered met in the current OT paradigm. Incorporating mission sets into an analysis provides a novel way of conducting operational testing when summary statistics are used to evaluate a measure. By marginalizing over the mission space, the Bayesian mission mean method takes into account the variability in performance among operational environments. In contrast, when using a grand mean, the small variance of ϕ_{GM} masks the complexity of the operational environment; therefore, decisions made based on the grand mean are more sensitive to outliers in performance among the mission sets.

Next, Chapter 3 broadened the work in phase II clinical trials to the OT paradigm and extended that work to a fully Bayesian method for a continuous response with an ANOVA structure—a structure that is common to OT. In doing so, we created a method using predictive probabilities (PP) that could be used in OT to end testing early. Incorporating PP into OT provides testers a way to make conclusions about a measure during an operational test and saving test runs. While constraints still exist for using predictive probabilities for interim analysis with continuous responses, the computational power has greatly increased since Geisser and W. Johnson (1994), and to a lesser degree since Dmitrienko and Wang (2006), which allowed for a fully Bayesian solution to this problem. Although the computational power has increased, the complexity of this computation, as acknowledged by Zhou et al. (2018), still remains. To both support efficient computations and limit further complexity than was already introduced by using an ANOVA structure for the continuous response, conjugate priors were used. Using a non-conjugate prior for η was also explored; however, the method became impractical for practitioners due to the extra computational power required. More efficient sampling methods may be possible that would allow for efficient sampling when using non-conjugate priors for a continuous response, and is an area for further research.

Finally, Chapter 4 introduced the novel conditional normalized partial borrowing power prior (CNPBPP) and provided a way to incorporate potentially dissimilar DT data into OT through use of CNPBPP. This new informative prior will give testers the ability to appropriately incorporate DT information that is currently not considered. The example used in this research demonstrated that it is possible that CNPBPP is less likely to borrow from DT information than NPBPP, which is a benefit for this area of research. This prior, based on conditional probability, is computationally more efficient than the previously proposed normalized partial borrowing power prior (NPBPP). Furthermore, by incorporating DT information, it is possible to get more precise estimates of model parameters, which can allow

for stopping OT earlier than when using reference priors. Finally, CNPBPP has protections in place to mitigate the risk of DT data creating a disinformative prior, as a_0 will always be allowed to be between 0 and 1.

While Tables 2.3, 3.1, 4.2, and 4.4 demonstrate the utility of the methods presented in this research, further work would be needed to evaluate type I and II error rates with a high degree of precision. Due to computational constraints, this was not explored; however, such work could involve a grid search to select θ_L and θ_U such that the error was controlled, as employed by Lee and D. D. Liu (2008). Another limitation with this method is the lack of access to DT data (National Research Council 1998). If DT data is unavailable, the methods presented in Chapters 2 and 3 can still be used, with the option of developing informative priors from subject matter expert (SME) opinion instead of reference priors (see Bedrick, R. Christensen, and W. Johnson 1996 for how this can be accomplished). Finally, as highlighted by Dickinson et al. (2015), the analysis employed within the current OT paradigm can be easily accomplished in Excel. To calculate the PPs in Chapters 3 and 4, a custom sampler was developed in R for the specific electric semi-truck example—a requirement that is considerably more complex than the current process. Even if an operational tester were familiar with R and Bayesian methods, an R package (or package in any software) would not currently be available to implement the two-stage sampling method used in Chapters 3 and 4 (whether DT is included or not). In order to mitigate this limitation to a degree, it is our intention to develop an R package to make the method more accessible to practitioners.

There are four main avenues we intend to pursue in future work. The first is to consider incorporating DT information that does not mirror OT. One extension would be to assume that DT models what OT considers to be latent or nuisance parameters, or that DT considers a different set of (related) parameters than OT. Another extension would be to consider a case where DT collects information on

a response variable that is related to a response variable in OT, but is not the OT response variable. Both of these extensions would be less restrictive than the method presented in Chapter 4.

Another avenue for future work reconsiders the treatment of factors and levels in the current OT paradigm. While also assumed in this research, it is unlikely that the factors affecting the response are independent (an assumption shared by the current OT paradigm); by designing an experiment for the selected factors and levels, and asking experts questions about the response for test runs (instead of the narrow focus of a single level), a better understanding of how these factors and levels create the mission space will be obtained. This can then be leveraged to develop a mission set that incorporates the complex nature of the operational environment of, e.g., Florida—rather than assuming independent factors and levels.

This research explored one option for adaptive testing—interim analysis. However, there are other adaptive methods that could also be incorporated. As addressed in Chapter 1, it is important to make efficient use of public funds. Chapters 3 and 4 provide a way to allow for cost savings, but this method can be extended to a fully adaptive test. Future work intends to explore how saved resources (either range time or physical test resources) can be re-allocated to other areas of testing that require more data.

Finally, as previously discussed, this method looks at the most granular evaluation within the OT framework—evaluating a single measure. Within the current OT paradigm, measures are grouped under critical operational issues (COIs), where COIs are developed from overarching themes that arise from the collections of requirements associated with a system. These COIs will ultimately be resolved as met or not met, and the collection of COI evaluations inform the decision at the system level (mission capable or not). Currently, this process is a subjective roll-up. In such a case, where decisions involve multiple measures, future work to address this

problem would involve the application of statistical decision theory. This could be accomplished by reducing the collection of measures into a univariate summary ϕ^* through a well-specified utility function expressing the importance of each measure to the overall decision of whether or not to approve the system. A threshold ϕ_0^* could then be specified for ϕ^*, and predictive probabilities could be obtained. With these lines of effort, it is possible to continue providing efficiencies within OT, making effective use of limited testing resources.

Appendix A

Overview of Bayesian Statistics

This appendix reviews pertinent concepts within the Bayesian framework for the methods presented in Chapters 2–4. The discussion in this appendix is derived from R. Christensen et al. (2011), where a more detailed discussion of these, and other, concepts can be found. The structure and form of the data to be collected, given the model parameter(s), is described through the data model (or density). The data model is a joint distribution between all observations within the data set, conditioned on the model parameters. Within the Bayesian framework, every parameter in a data model is unknown and has a prior distribution associated with it that is developed independently of the data to be collected. This prior distribution is used to quantify the uncertainty surrounding a given model parameter and represents an individual's beliefs about the model parameter. Therefore, priors are a means of incorporating key information about the model parameter into a statistical analysis. For example, consider binomial the data model $\text{Bin}(n, \theta)$; a natural prior choice for θ would be a $\text{Beta}(a, b)$ distribution, where a and b are selected such that the distribution reflects prior beliefs about θ. A more detailed discussion of priors will be given in Section 4.1.

Given the data model, prior distribution(s), and data, Bayes' Theorem can be used to create a posterior distribution. The posterior distribution is foundational for Bayesian inference, representing an updated belief about parameters based on the observed data. Using Bayes' Theorem, the posterior distribution is:

$$\begin{aligned} p(\theta|Y) &= \frac{f(Y|\theta)p(\theta)}{\int f(Y|\theta)p(\theta)d\theta} \\ &= \frac{f(Y|\theta)p(\theta)}{f(Y)} \\ &\propto f(Y|\theta)p(\theta). \end{aligned}$$

Often, the calculus to obtain a posterior distribution is intractable; therefore, numerical approximations, such as *Markov chain Monte Carlo* (MCMC), are often used in practice. "The idea of Markov chain Monte Carlo is to define a sequence of random vectors $\theta^1, \theta^2, \theta^3, \ldots$ in which the distribution of θ^k near the beginning of the sequence can be just about anything but in which the distribution will eventually settle down to the posterior distribution" (BIDA). Therefore, once the Markov chain converges to the posterior distribution, it will remain in the posterior distribution due to stationarity. Under some mild conditions and with a sufficiently large k, the θ^k's from a MCMC sampler will come from the posterior distribution. These θ^k (also referred to as samples, posterior draws, or MCMC draws) can then be used to numerically approximate the posterior distribution, and estimate functionals of that distribution (such as expectations, quantiles, etc).

Two common MCMC methods are a *Gibbs sampler* and a *Metropolis-within-Gibbs sampler*. "Gibbs sampling is a method for constructing a Markov chain that is extremely useful when one can isolate the conditional distribution of each parameter given all the other parameters", which is referred to as a *full conditional* (BIDA). Fundamentally, a Gibbs sampler will iteratively sample from each full conditional to wander into the posterior distribution. When these full conditionals are recognizable distributions that are easy to sample from (e.g. normal distribution, beta distri-

bution), a Gibbs sampler will be computationally efficient. When one or more of the full conditionals are not recognizable, a more computationally inefficient method must be used. A Metropolis-within-Gibbs sampler is one such method. Metropolis-within-Gibbs is a "hybrid sampler that replaces a sample from a full conditional [in a Gibbs sampler] with one step of the Metropolis algorithm" when the full conditional is not recognizable (BIDA). This replacement step is more computationally complex than sampling from a known distribution, and ultimately runs slower than a Gibbs sampler (i.e. is more computationally inefficient).

Once a posterior distribution is obtained, various estimates of interest (e.g., point or interval estimate for a parameter) can be calculated. For instance, the posterior mean is

$$E(\theta|y) = \int \theta p(\theta|y) d\theta.$$

Probabilities such as $\Pr(\theta > c|Y)$, where c is a constant, can also be calculated:

$$\Pr(\theta > c|Y) = \int_c^\infty p(\theta|Y) d\theta.$$

When the posterior distribution is obtained through numerical approximations, the posterior samples can be used to numerically approximate the estimates of interest. For example, the posterior mean can be numerically approximated by taking the mean of the posterior samples (after removing the *burn-in samples*).

As alluded to previously, when using MCMC methods, there is no guarantee that the chain will begin in the posterior distribution; instead, it may take time before it converges to the posterior (as, subject to some common regularity conditions, it is guaranteed to eventually do). Therefore, the samples obtained before converging to the posterior distribution (referred to as burn-in samples) are removed, ensuring that all posterior inferences are made using samples from the posterior distribution. The number of burn-in samples can be determined visually by looking at trace plots and noting where convergence seems to have occurred.

Finally, Bayesian inference allows for the calculation of predictive distributions and predictive probabilities. The predictive distribution for a future observation(s), $\tilde{y}$, given the past observations is:

$$f_p(\tilde{y}|Y) = \int f_p(\tilde{y}|\theta)p(\theta|Y)d\theta.$$

Using the predictive distribution, estimates of interest can be found in the same manner described for the posterior distribution. An advantage of the predictive density is that one can calculate the predictive probability of various estimates of interest. For example, if the interest is in the probability that the next observation will be less than or equal to 5 given what has been observed, the predictive probability is:

$$\Pr[\tilde{y} \leq 5|Y] = \int_{-\infty}^{5} f_p(\tilde{y}|Y)d\tilde{y}.$$

Instead of considering $\Pr[\tilde{y} \leq 5|y]$, one could consider $\Pr[\tilde{y} \leq 5|\theta]$—the probability that the next observation will be less than or equal to 5, given the parameter. Christensen, et al. (2011) show that the posterior mean of the probability that $\tilde{y} \leq 5$ is the same as this predictive probability. That is to say,

$$E[\Pr(\tilde{y} \leq 5|\theta)|y] \equiv \Pr[\tilde{y} \leq 5|y] = \int_{-\infty}^{5} f_p(\tilde{y}|y)d\tilde{y}.$$

A more detailed discussion of predictive probability will be given in Chapter 3.

Appendix B

Method for Generating Simulated Operational Testing Data

This appendix details the simulation method used to generate the OT data for the electric semi-truck example in Chapters 2–4. The simulation method first generates group means; then applies a transformation to standard normal errors to obtain new errors; and, finally, creates the responses by adding the transformed error to the group means.

To present this simulation method, the following outlines how data set 4 was generated. Let X be the randomized design matrix for a 2^4 full factorial experimental design with five replicates; the columns of X relate to model parameters and the rows correspond to test events. The rows of X are indexed by i, $i = 1, \ldots, 80$; when i is used in other notation, it represents that the notation corresponds to the i^{th} row of X. Let β be a 13×1 column vector of the true model parameter values (as defined in Table 2.4): $\beta' = [349, 50, 15, 0, -5, 75, 50, 10, 50, 25, 50, 25, 25]$. Let μ be a 80×1 column vector of true group means associated with each test event, and let μ_i be the i^{th} element of μ. Then, μ is obtained through $\mu = X\beta$.

Next, standard errors are generated and then transformed. Let ε be a 80×1 column vector of standard normal errors, where ε_i represents the i^{th} element of ε. It was reasonable to assume that a mean-variance relationship would exist for electric semi-truck travel distances on a fixed amount of charge; additionally, this relationship could be used to determine how the method worked when data weren't perfectly generated for the evaluation model. To accomplish this, let $s = 2$ and a be the following column vector: $a' = [0, s^2, s, 0, 0, s, s^2, 0, 0, 0, 0, 0, 0]$. The mean-variance relationship is then created by obtaining the column vector Xa. Furthermore, let $b = 50$, which provides a baseline change for this transformation of ε, and let $J_{80}^1 b$ be a 80×1 column vector containing the value b in every row. Let ϵ represent the transformed ε and let ϵ_i be the i^{th} element of ϵ. The transformation (using element-wise matrix multiplication) is then applied to ε in the following manner: $\epsilon = \varepsilon \circ (J_{80}^1 b + Xa)$. Finally, the i^{th} response for data set 4, y_i, is defined by $y_i = \mu_i + \epsilon_i$.

The simulation method presented above was used to generate each data set, using different η values and different transformations of ε. For comparison purposes, the same X and ε were used for each data set. In this paper, data sets were developed from a combination of seven different ηs and three different transformations of ε (defined by the values s and b take on). The η value used in a given data set can be found in Table 3.1. Transformation 1 (data sets 1 – 7) used $s = 2$ and $b = 50$; transformation 2 (data sets 8 - 14) used $s = 3$ and $b = 100$; and, finally, transformation 3 (data sets 15 - 21) used $s = 4$ and $b = 150$. These transformations create three different variance structures (or, alternatively, precision structures) for the data sets, while the different values of η shift distribution of the data.

In Section 4.3, not all observations were seen when calculating PP. To mimic what the practitioner would see during test execution, the observations yet to be seen were removed from the data set. For example, when $n_0 = 75$, the last five

observations of each data set ($y_i, i = 76, \ldots, 80$ for each data set) were removed. As described in the paper, the remaining values to be seen (e.g. the last five that had now been removed from the data set) would then be imputed by the outside sampler to implement the method.

Appendix C

Full Conditionals using Normalized Partial Borrowing Power Priors

This appendix contains the derivations for the full conditionals presented Section 4.1.7. The full conditional for $\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \tau, a_0, D_0, D$ is:

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \tau, a_0, D_0, D) \\
&\quad \propto \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)\Big)\bigg) \\
&\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0)\Big)\bigg)
\end{aligned}
$$

Using the Proof from Theorem 2.2.1 in R. Christensen (2020):

$$
\begin{aligned}
&= \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)\Big)\bigg) \\
&\qquad \times \exp\bigg(-\frac{\tau}{2}\Big((MY - Y)'(MY - Y)\Big)\bigg) \\
&\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)\bigg) \\
&\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((M_0Y_0 - Y_0)'(M_0Y_0 - Y_0)\Big)\bigg)
\end{aligned}
$$

where $M = X(X'X)^{-1}X'$ and $M_0 = X_0(X_0'X_0)^{-1}X_0'$

$$\begin{aligned}
&\propto \exp\bigg(-\frac{\tau}{2}\bigg((X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\bigg((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\bigg)\bigg) \\
&= \exp\bigg(-\frac{\tau}{2}\bigg((X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})'\Sigma_0^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})\bigg)
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{0h} = (X_0'X_0)^{-1}X_0'Y_0$, and $\Sigma_0^{-1} =_0' X_0$

Using results from Appendix D,

$$\begin{aligned}
&= \exp\bigg(-\frac{\tau}{2}\bigg((X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) \\
&\qquad + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})'\Sigma_0^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})\bigg)
\end{aligned}$$

where $M_{22} = X_2(X_2'X_2)^{-1}X_2'$

$$\begin{aligned}
&\propto \exp\bigg(-\frac{\tau}{2}(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})'\Sigma_0^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})\bigg) \\
&= \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0)'\Sigma_2^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})'\Sigma_0^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})\bigg)
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_0 = (X_2'X_2)^{-1}X_2'Y - (X_2'X_2)^{-1}X_2'X_1\boldsymbol{\beta}_1$ and let $\Sigma_2^{-1} = \tau(X_2'X_2)$

$$= \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0)'\Sigma_2^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_0) + (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})'\Sigma_0^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{0h})\bigg)$$

Using Proposition 2 (Sum of two quadratic forms in x) from Rosenberg n.d.:

$$= \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_0 - O_0^{-1}b_0)'O_0(\boldsymbol{\beta}_0 - O_0^{-1}b_0) - b_0'O_0^{-1}b_0 + R_0)\bigg)$$

where $O_0 = \tau\bigg((X_2'X_2) + a_0(X_0'X_0)\bigg)$, $\quad b_0 = \tau\bigg((X_2'X_2)\hat{\boldsymbol{\beta}}_0 + a_0(X_0'X_0)\hat{\boldsymbol{\beta}}_{0h}\bigg)$,

and $\quad R_0 = \tau\bigg(\hat{\boldsymbol{\beta}}_0'(X_2'X_2)\hat{\boldsymbol{\beta}}_0 + a_0\hat{\boldsymbol{\beta}}_{0h}'(X_0'X_0)\hat{\boldsymbol{\beta}}_{0h}\bigg)$

$$\propto \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_0 - O_0^{-1}b_0)'(O_0^{-1})^{-1}(\boldsymbol{\beta}_0 - O_0^{-1}b_0)\bigg)\bigg)$$

This the kernel of a normal distribution. Then, $\boldsymbol{\beta}_0|\tau, a_0, D_0 \sim \mathcal{N}(\boldsymbol{\mu}_{0f}, \tau^{-1}\Lambda_{0f}^{-1})$, where

$\boldsymbol{\Lambda}_{0f} = (X_2'X_2) + a_0(X_0'X_0)$ and $\boldsymbol{\mu}_{0f} = O_0^{-1}b_0 = \boldsymbol{\Lambda}_{0f}^{-1}\bigg(X_2'Y - X_2'X_1\boldsymbol{\beta}_1 + a_0X_0'Y_0\bigg)$.

Next, the full conditional for $\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_0, D_0, D$ is:

$$\begin{aligned} &p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_0, D_0, D) \\ &\quad\propto \exp\bigg(-\frac{\tau}{2}\bigg((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)\bigg)\bigg) \\ &\qquad\times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg) \end{aligned}$$

Using the Proof from Theorem 2.2.1 in R. Christensen (2020):

$$\begin{aligned} &= \exp\bigg(-\frac{\tau}{2}\bigg((X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)\bigg)\bigg) \\ &\qquad\times \exp\bigg(-\frac{\tau}{2}\bigg((MY - Y)'(MY - Y)\bigg)\bigg) \\ &\qquad\times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg) \\ &\propto \exp\bigg(-\frac{\tau}{2}\bigg((X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)\bigg)\bigg) \end{aligned}$$

$$\times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg)$$

Using the same process as in Appendix D, but using the perpendicular projection operator onto $C(X_1)$ instead,

$$\begin{aligned}
&= \exp\bigg(-\frac{\tau}{2}\bigg((X_1\boldsymbol{\beta}_1 + M_1X_2\boldsymbol{\beta}_0 - M_1Y)'(X_1\boldsymbol{\beta}_1 + M_1X_2\boldsymbol{\beta}_0 - M_1Y) \\
&\qquad + ((I - M_1)X_2\boldsymbol{\beta}_0 - M_2Y)'((I - M_1)X_2\boldsymbol{\beta}_0 - M_2Y)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg)
\end{aligned}$$

where $M_1 = X_1'(X_1'X_1)^{-1}X_1'$

$$\begin{aligned}
&\propto \exp\bigg(-\frac{\tau}{2}\bigg((X_1\boldsymbol{\beta}_1 + M_1X_2\boldsymbol{\beta}_0 - M_1Y)'(X_1\boldsymbol{\beta}_1 + M_1X_2\boldsymbol{\beta}_0 - M_1Y)\bigg)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg) \\
&= \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'\Sigma_1^{-1}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)\bigg) \\
&\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg)
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'Y - (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_0$ and let $\Sigma_1^{-1} = \tau(X_1'X_1)$

$$= \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)'\Sigma_1^{-1}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1) + (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\tau\boldsymbol{\Lambda}_1)(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\bigg)\bigg)$$

Again using Proposition 2 (Sum of two quadratic forms in x) from Rosenberg (n.d.):

$$= \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_1 - O_1^{-1}b_1)'O_1(\boldsymbol{\beta}_1 - O_1^{-1}b_1) - b_1'O_1^{-1}b_1 + R_1)\bigg)$$

where $O_1 = \tau\Big(X_1'X_1 + \boldsymbol{\Lambda}_1\Big)$, $\quad b_1 = \tau\Big((X_1'X_1)\hat{\boldsymbol{\beta}}_1 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big)$,

and $\quad R_1 = \tau\Big(\hat{\boldsymbol{\beta}}_1'(X_1'X_1)\hat{\boldsymbol{\beta}}_1 + \boldsymbol{\mu}_1'\boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big)$,

$$\propto \exp\bigg(-\frac{1}{2}\bigg((\boldsymbol{\beta}_1 - O_1^{-1}b_1)'(O_1^{-1})^{-1}(\boldsymbol{\beta}_1 - O_1^{-1}b_1)\bigg)$$

This is the kernel of a normal distribution: $\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_1, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{1f}, \tau^{-1}\boldsymbol{\Lambda}_{1f}^{-1})$, where $\boldsymbol{\mu}_{1f} = \boldsymbol{\Lambda}_{1f}^{-1}\Big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big)$ and $\boldsymbol{\Lambda}_{1f} = (X_1'X_1) + \boldsymbol{\Lambda}_1$.

Next, the full conditional for $\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D$ is:

$$
\begin{aligned}
&p(\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D) \\
&\quad\propto (\tau)^{\frac{n}{2}}(\tau)^{(\frac{n_0a_0}{2}+\alpha_0)-1}\exp(-\gamma_0\tau)\exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg) \\
&\qquad\qquad\times\exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0-Y_0)\Big)\bigg) \\
&\qquad\qquad\times(\tau)^{\frac{p_1}{2}}\exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\bigg) \\
&\quad= (\tau)^{(\frac{n+n_0a_0+p_1}{2}+\alpha_0)-1}\exp(-\gamma_0\tau)\exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg) \\
&\qquad\qquad\times\exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0-Y_0)\Big)\bigg) \\
&\qquad\qquad\times\exp\bigg(-\tau\Big(\frac{1}{2}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\Big)\bigg) \\
&\quad= (\tau)^{(\frac{n+n_0a_0}{2}+\alpha_0)-1}\exp\bigg(-\tau\Big(\frac{1}{2}\Big[(X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y) \\
&\qquad\qquad+a_0(X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0-Y_0)+(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\Big] \\
&\qquad\qquad+\gamma_0\Big)\bigg)
\end{aligned}
$$

Let $\alpha_f = \frac{n+n_0a_0+p_1}{2} + \alpha_0$ and

$\gamma_f = \frac{1}{2}[(X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)+a_0(X_0\boldsymbol{\beta}_0-Y_0)'(X_0\boldsymbol{\beta}_0-Y_0)+(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1)(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)]+\gamma_0$

$$= (\tau)^{\alpha_f-1}\exp(-\gamma_f\tau)$$

This is the kernel of an inverse gamma distribution. Therefore,
$\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D \sim \text{Gam}(\alpha_f, \gamma_f)$.

Finally, the full conditional for $a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D$:

$$
\begin{aligned}
&p(a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D) \\
&\quad \propto (a_0)^{\frac{p_0}{2}} \frac{\gamma_{0n}^{\alpha_{0n}}}{\Gamma(\alpha_{0n})} (\tau)^{(\frac{n_0 a_0}{2} + \alpha_0) - 1} I_A(a_0) \\
&\qquad\qquad \times \exp\left(- \frac{a_0 \tau}{2} \left((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0) \right) \right)
\end{aligned}
$$

Appendix D

Separating Out $\boldsymbol{\beta}_0$

This appendix contains the derivations for separating $(X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY)$ into two terms—one term that is independent of $\boldsymbol{\beta}_0$ and one term that depends on $\boldsymbol{\beta}_0$. The results form this appendix are then used in Appendix C to obtain the full conditionals when using a normalized partial borrowing power prior.

$$\begin{aligned}(X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY) &= \left([X_1, X_2]\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_0\end{bmatrix} - MY\right)'\left([X_1, X_2]\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_0\end{bmatrix} - MY\right)\\ &= (X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_0 - MY)'(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_0 - MY)\\ &= (X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY)'(X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY)\end{aligned}$$

First, consider $(X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY)$. Let $C(\cdot)$ represent a column space. M, the ppo onto the $C(X)$, can then also be partitioned. Consider that $C(X) = C(X_1, X_2)$. Let M_{22} be the ppo onto the $C(X_2)$; that is, $M_{22} = X_2(X_2'X_2)^{-1}X_2'$. Consider the following, as shown in R. Christensen (2020):

$$\begin{aligned}X_1 &= IX_1\\ &= (I - M_{22} + M_{22})X_1\end{aligned}$$

$$= (I - M_{22})X_1 + M_{22}X_1$$

Then, $C(X) = C((I - M_{22})X_1, M_{22}X_1, X_2)$. Since M_{22} projects into $C(X_2)$, $C(M_{22}X_1, X_2) = C(X_2)$. Then, $C(X) = C((I - M_{22})X_1, X_2)$. M is then the ppo onto $C(X) = C((I - M_{22})X_1, X_2)$, with $(I - M_{22})X_1$ and X_2 orthogonal. Therefore, by Theorem B.45, M is the sum of the ppos for subspaces $C((I-M_{22})X_1)$ and $C(X_2)$ (R. Christensen 2020):

$$\begin{aligned} M &= (I - M_{22})X_1(X_1'(I - M_{22})X_1)^{-1}X_1'(I - M_{22}) + M_{22} \\ &= M_{21} + M_{22} \end{aligned}$$

Then,

$$\begin{aligned} &(X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY) \\ &\quad = (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 + (I - M_{22})X_1\boldsymbol{\beta}_1 - MY) \\ &\quad = (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 + (I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y - M_{22}Y) \\ &\quad = ((X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)) \\ &\quad = ((X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)) \end{aligned}$$

Let $A = (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)$ and $B = ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)$. Then,

$$= (A + B)$$

Similarly, $(X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY)' = (A + B)' = (A' + B')$. Therefore,

$$
\begin{aligned}
&(X\boldsymbol{\beta}-MY)'(X\boldsymbol{\beta}-MY) \\
&\quad = (X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY)'(X_2\boldsymbol{\beta}_0 + X_1\boldsymbol{\beta}_1 - MY) \\
&\quad = (A' + B')(A + B) \\
&\quad = (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) \\
&\qquad\quad + (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y) \\
&\qquad\quad + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) \\
&\qquad\quad + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)
\end{aligned}
$$

However, the cross-product terms are 0:

$$
\begin{aligned}
&(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y) \\
&\quad = ((X_2\boldsymbol{\beta}_0)' + (M_{22}X_1\boldsymbol{\beta}_1)' - (M_{22}Y)')((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y) \\
&\quad = (X_2\boldsymbol{\beta}_0)'(I - M_{22})X_1\boldsymbol{\beta}_1 - (X_2\boldsymbol{\beta}_0)'M_{21}Y \\
&\qquad\quad + (M_{22}X_1\boldsymbol{\beta}_1)'(I - M_{22})X_1\boldsymbol{\beta}_1 - (M_{22}X_1\boldsymbol{\beta}_1)'M_{21}Y \\
&\qquad\quad - (M_{22}Y)'(I - M_{22})X_1\boldsymbol{\beta}_1 + (M_{22}Y)'M_{21}Y)
\end{aligned}
$$

M_{21} and M_{22} are ppos, so they are idempotent and symmetric. $(I - M_{22})$ is also a ppo, so idempotent and symmetric.

$$
\begin{aligned}
&= \boldsymbol{\beta}_0'X_2'(I - M_{22})X_1\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0'X_2'M_{21}Y + \boldsymbol{\beta}_1'X_1'M_{22}(I - M_{22})X_1\boldsymbol{\beta}_1 \\
&\qquad\quad - \boldsymbol{\beta}_1'X_1'M_{22}M_{21}Y - Y'M_{22}(I - M_{22})X_1\boldsymbol{\beta}_1 + Y'M_{22}M_{21}Y)
\end{aligned}
$$

Because $M_{22} \perp (I - M_{22})$ and $M_{22} \perp M_{21}$, that means $M_{22}(I - M_{22}) = 0$ and $M_{22}M_{21} = 0$. Also, $X_2'(I - M_{22}) = 0$, which also means $X_2'M_{21} = 0$, by definition.

$$
= 0
$$

Appendix D. Separating Out $\boldsymbol{\beta}_0$

Similarly,

$$\begin{aligned}((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) \\ = ((X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y))' \\ = 0\end{aligned}$$

Therefore,

$$\begin{aligned}(X\boldsymbol{\beta} - MY)'(X\boldsymbol{\beta} - MY) \\ = (X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y)'(X_2\boldsymbol{\beta}_0 + M_{22}X_1\boldsymbol{\beta}_1 - M_{22}Y) \\ + ((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)'((I - M_{22})X_1\boldsymbol{\beta}_1 - M_{21}Y)\end{aligned}$$

Appendix E

Full Conditionals using Conditional Normalized Partial Borrowing Power Priors

This appendix contains the derivations for the full conditionals presented Section 4.2. Beginning with the full conditional for $\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \tau, a_0, D_0, D$:

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1,\tau, a_0, D_0, D) \\
&\qquad \propto \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y)\Big)\bigg) \\
&\qquad\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - Y_0)'(X_0\boldsymbol{\beta}_0 - Y_0)\Big)\bigg)
\end{aligned}
$$

This is the same as in Appendix C; then, $\boldsymbol{\beta}_0|\tau, a_0, D_0 \sim \mathcal{N}(\boldsymbol{\mu}_{0c}, \tau^{-1}\Lambda_{0c}^{-1})$, where $\boldsymbol{\Lambda}_{0c} = (X_2'X_2) + a_0(X_0'X_0)$ and $\boldsymbol{\mu}_{0c} = O_0^{-1}b_0 = \boldsymbol{\Lambda}_{0c}^{-1}\Big(X_2'Y - X_2'X_1\boldsymbol{\beta}_1 + a_0X_0'Y_0\Big)$.

Next, the full conditional for $\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_0, D_0, D$:

$$\begin{aligned} p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0,&\tau, a_0, D_0, D) \\ &\propto \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg) \\ &\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\bigg) \end{aligned}$$

This is the same as in Appendix C; then, $\boldsymbol{\beta}_1|\boldsymbol{\beta}_0, \tau, a_1, D_0, D \sim \mathcal{N}(\boldsymbol{\mu}_{1c}, \tau^{-1}\boldsymbol{\Lambda}_{1c}^{-1})$, where $\boldsymbol{\mu}_{1c} = \boldsymbol{\Lambda}_{1c}^{-1}\Big(X_1'Y - X_1'X_2\boldsymbol{\beta}_0 + \boldsymbol{\Lambda}_1\boldsymbol{\mu}_1\Big)$ and $\boldsymbol{\Lambda}_{1c} = (X_1'X_1) + \boldsymbol{\Lambda}_1$.

Next, the full conditional for $\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D$:

$$\begin{aligned} p(\tau|\boldsymbol{\beta}_0,&\boldsymbol{\beta}_1, a_0, D_0, D) \\ &\propto (\tau)^{\frac{n}{2}} \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg)(\tau)^{\frac{p_0}{2}} \\ &\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0-M_0Y_0)'(X_0\boldsymbol{\beta}_0-M_0Y_0)\Big)\bigg)(\tau)^{\frac{p_1}{2}} \\ &\qquad \times \exp\bigg(-\frac{1}{2}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\tau^{-1}\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\bigg) \\ &\qquad \times (\tau)^{\alpha_0-1}\exp(-\gamma_0\tau) \end{aligned}$$

Recall: $p = p_0 + p_1$.

$$\begin{aligned} &= (\tau)^{\frac{n+p}{2}+\alpha_0-1} \exp\bigg(-\frac{\tau}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg) \\ &\qquad \times \exp\bigg(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0-M_0Y_0)'(X_0\boldsymbol{\beta}_0-M_0Y_0)\Big)\bigg) \\ &\qquad \times \exp\bigg(-\frac{\tau}{2}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1-\boldsymbol{\mu}_1)\bigg)\exp(-\gamma_0\tau) \\ &= (\tau)^{\frac{n+p}{2}+\alpha_0-1} \exp\bigg(-\tau\bigg[\frac{1}{2}\Big((X\boldsymbol{\beta}-Y)'(X\boldsymbol{\beta}-Y)\Big)\bigg) \end{aligned}$$

$$+ a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)$$
$$+ (\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1^{-1})^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\Big) + \gamma_0\Big]\Big)$$

Let and $\alpha_c = \frac{1}{2}(n+p) + \alpha_0$ and

$$\gamma_c = \tfrac{1}{2}\Big((X\boldsymbol{\beta} - Y)'(X\boldsymbol{\beta} - Y) + a_0(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0) +$$

$$(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)'(\boldsymbol{\Lambda}_1)(\boldsymbol{\beta}_1 - \boldsymbol{\mu}_1)\Big) + \gamma_0$$

$$= (\tau)^{\alpha_c - 1}\exp(-\gamma_c\tau)$$

This is the kernel of an inverse gamma distribution, using the shape and rate parameterization. Therefore, $\tau|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, a_0, D_0, D \sim \text{Gam}(\alpha_f, \gamma_f)$.

Finally, the full conditional for $a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D$:

$$p(a_0|\boldsymbol{\beta}_0,\boldsymbol{\beta}_1, \tau, D_0, D)$$
$$\propto a_0^{\frac{p_0}{2}}\exp\Big(-\frac{a_0\tau}{2}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)\Big)I_A(a_0)$$
$$= a_0^{(\frac{p_0}{2}+1)-1}\exp\Big(-a_0\Big(\frac{\tau}{2}(X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)\Big)I_A(a_0)$$

This is the kernel of a truncated gamma distribution (TG), with the shape and rate parameterization. The R package "TruncatedDistributions" can sample from a TG, but uses the shape / scale parameterization. Note, however, that the scale is equal to 1/rate. Therefore, using the shape / scale parameterization

$$a_0|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau, D_0, D \sim TG\Big(\frac{p_0}{2}+1,\ \ 2\tau^{-1}\Big((X_0\boldsymbol{\beta}_0 - M_0Y_0)'(X_0\boldsymbol{\beta}_0 - M_0Y_0)\Big)^{-1},\ \ 0,\ \ 1\Big)$$

References

Barker, A. D., C. C. Sigman, G. J. Kelloff, N. M. Hylton, Donald A. Berry, and L. J. Esserman. "I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy". In: *Clinical Pharmacology and Therapeutics* 86.1 (2009), pp. 97–100. ISSN: 00099236. DOI: 10.1038/clpt.2009.68.

Bedrick, Edward J, Ronald Christensen, and Wesley Johnson. "A New Perspective on Priors for Generalized Linear Models". In: *Journal of American Statistical Association* 91.436 (1996), pp. 1450–1460.

Berry, Donald A. "A Case for Bayesianism in Clinical Trials". In: *Statistics in Medicine* 12.15-16 (1993), pp. 1377–1393. ISSN: 10970258. DOI: 10.1002/sim.4780121504.

— "A guide to drug discovery: Bayesian clinical trials". In: *Nature Reviews Drug Discovery* 5.1 (2006), pp. 27–36. ISSN: 14741776. DOI: 10.1038/nrd1927.

— "Interim Analysis in Clinical Trials : The Role of the Likelihood". In: 41.2 (1987), pp. 117–122.

— "Introduction to Bayesian methods III: Use and interpretation of Bayesian tools in design and analysis". In: *Clinical Trials* 2.4 (2005), pp. 295–300. ISSN: 17407745. DOI: 10.1191/1740774505cn100oa.

Berry, Scott M., Bradley P. Carlin, J. Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press, 2011, p. 305. ISBN: 9781439825488.

Boyd, John. *A Discourse on Winning and Losing*. Ed. by Grant T Hammond. Maxwell AFB: Air University Press, 2018. ISBN: 2017039845. URL: http://www.airuniversity.af.mil/AUPress/.

Chen, Ming-Hui, Joseph G. Ibrahim, H. Amy Xia, Thomas Liu, and Violeta Hennessey. "Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program". In: *Statistics in Medicine* 33.9 (2014), pp. 1600–1618. ISSN: 10970258. DOI: 10.1002/sim.6067.

Chen, Ming-Hui, Joseph G. Ibrahim, Peter Lam, Alan Yu, and Yuanye Zhang. "Bayesian Design of Noninferiority Trials for Medical Devices Using Historical Data". In: *Biometrics* 67.3 (2011), pp. 1163–1170. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2011.01561.x.

Christensen, Ronald. *Plane Answers to Complex Questions*. 5th. New York: Springer, 2020. DOI: https://doi.org/10.1007/978-1-4419-9816-3.

Christensen, Ronald, Wesley Johnson, Adam Branscum, and Timothy Hanson. *Bayesian Ideas and Data Analysis*. Boca Raton, FL: CRC Press, 2011. ISBN: 978-1-4398-0354-7.

Department of the Navy. *Operational Test Director's Manual (COMOPTEVFOR Instruction 3980.2I)*. Tech. rep. OPTEVFOR, 2019.

Dewald, Lee, Robert Holcomb, Sam Parry, and Alyson G. Wilson. "A Bayesian Approach to Evaluation of Operational Testing of Land Warfare Systems". In: *Military Operations Research* 21.4 (2016), pp. 23–32. DOI: 10.5711/1082598321423.

Dickinson, Rebecca M., Laura J. Freeman, Bruce A. Simpson, and Alyson G. Wilson. "Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study". In: *Journal of Quality Technology* 47.4 (2015), pp. 400–415.

Director Operational Test and Evaluation. *Guidance on the use of Design of Experiments in Operational Test and Evaluation*. Tech. rep. Washington, D.C., 2010, pp. 1–2.

Dmitrienko, Alexei and Ming Dauh Wang. "Bayesian predictive approach to interim monitoring in clinical trials". In: *Statistics in Medicine* 25.13 (2006), pp. 2178–2195. ISSN: 02776715. DOI: 10.1002/sim.2204.

Duan, Yuyan, Keying Ye, and Eric P. Smith. "Evaluating water quality using power priors to incorporate historical information". In: *Environmetrics* 17.1 (2006), pp. 95–106. ISSN: 11804009. DOI: 10.1002/env.752.

Encyclopaedia Britannica. *Beauford scale*. 2017. URL: https://www.britannica.com/science/Beaufort-scale.

Freeman, Laura J. and Catherine Warner. "Informing the Warfighter—Why Statistical Methods Matter in Defense Testing". In: *CHANCE* 31.2 (Apr. 2018), pp. 4–11. ISSN: 0933-2480. DOI: 10.1080/09332480.2018.1467627. URL: https://www.tandfonline.com/doi/full/10.1080/09332480.2018.1467627.

Geisser, Seymour and Wesley Johnson. "Interim analysis for normally distributed observables". In: *Multivariate Analysis and Its Applications.* 1994, pp. 263–279. DOI: 10.1214/lnms/1215463801.

Gelman, Andrew, John B. Carlin, Hall S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. 3rd Editio. CRC Press, 2014.

Hobbs, Brian P., Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. "Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials". In: *Biometrics* 67.3 (Sept. 2011), pp. 1047–1056. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2011.01564.x. URL: http://doi.wiley.com/10.1111/j.1541-0420.2011.01564.x.

Ibrahim, Joseph G. and Ming-Hui Chen. "Power prior distributions for regression models". In: *Statistical Science* 15.1 (2000), pp. 46–60. ISSN: 08834237. DOI: 10.1214/ss/1009212673.

Ibrahim, Joseph G., Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. "The power prior: Theory and applications". In: *Statistics in Medicine* 34.28 (Dec. 2015), pp. 3724–3749. ISSN: 10970258. DOI: 10.1002/sim.6728.

Ibrahim, Joseph G., Ming-Hui Chen, and Debajyoti Sinha. "On optimality properties of the power prior". In: *Journal of the American Statistical Association* 98.461 (2003), pp. 204–213. ISSN: 01621459. DOI: 10.1198/016214503388619229.

Ibrahim, Joseph G., Ming-Hui Chen, H. Amy Xia, and Thomas Liu. "Bayesian Meta-Experimental Design: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes". In: *Biometrics* 68.2 (2012), pp. 578–586. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2011.01679.x.

Johnson, Rachel T., Gregory T. Hutto, James R. Simpson, and Douglas C. Montgomery. "Designed Experiments for the Defense Community". In: *Quality Engineering* 24.1 (2012), pp. 60–79. ISSN: 08982112. DOI: 10.1080/08982112.2012.627288.

Johnstone, Iain M. "High dimensional Bernstein-von Mises: simple examples". In: 6 (2010), pp. 87–98. ISSN: 1939-4039. DOI: 10.1214/10-imscoll607.

Joint Chiefs of Staff. *Manual for the Operation of the Joint Capabilities Integration and Development System (JCIDS Manual)*. 2018. URL: https://www.dau.edu/cop/rqmt/DAU%20Sponsored%20Documents/Manual%20-%20JCIDS,%2031%20Aug%202018.pdf.

Kendall, Frank. *Department of Defense Instruction 5000.02T: Operation of the Defense Acquisition System*. Tech. rep. 2015. URL: http://www.dtic.mil/whs/directives/corres/pdf/850001%7B%5C_%7D2014.pdf.

Lee, J. Jack and Diane D. Liu. "A predictive probability design for phase II cancer clinical trials". In: *Clinical Trials* 5.2 (2008), pp. 93–106. ISSN: 17407745. DOI: 10.1177/1740774508089279.

Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and Jim O. Berger. "Mixtures of g priors for Bayesian variable selection". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 410–423. ISSN: 01621459. DOI: 10.1198/016214507000001337.

Liu, Meng and Emily V. Dressler. "A predictive probability interim design for phase II clinical trials with continuous endpoints". In: *Statistics in Medicine* 37.12 (2018), pp. 1960–1972. ISSN: 10970258. DOI: 10.1002/sim.7659.

Mattis, Jim. *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competative Edge*. Tech. rep. Washington, D.C.: U.S. Department of Defense, 2018. URL: https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

Montgomery, Douglas C. *Design and Analysis of Experiments*. 8th. New York: Wiley, 2012. ISBN: 978-1118-14692-7.

Nadarajah, Saralees and Samuel Kotz. "R Programs for Computing Truncated Distributions". In: *Journal of Statistical Software* 16.Code Snippet 2 (2006), pp. 1–8. ISSN: 1548-7660. DOI: 10.18637/jss.v016.c02.

National Research Council. *Improved Operational Test and Evaluation Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report*. Washington, D.C.: National Academy Press, 2004.

— *Statistical Issues in Defense Analysis and Testing : Summary of a Workshop*. Ed. by John E Rolph and Duane L Steffey. Washington, D.C.: National Academy Press, 1994. DOI: 10.17226/9686. URL: https://search-ebscohost-com.libproxy.unm.edu/login.aspx?direct=true%7B%5C&%7Ddb=nlebk%7B%5C&%7DAN=123732%7B%5C&%7Dsite=eds-live%7B%5C&%7Dscope=site.

— *Statistics, Testing, and Defense Acquisition. : New Approaches and Methodological Improvements.* Ed. by Michael L Cohen, John E Rolph, and Duane L Steffey. Washington, D.C.: National Academy Press, 1998. URL: http://libproxy.unm.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%7B%5C&%7Ddb=cat06111a%7B%5C&%7DAN=unm.EBC3375607%7B%5C&%7Dsite=eds-live%7B%5C&%7Dscope=site%20https://ebookcentral.proquest.com/lib/unm/detail.action?docID=3375607.

Neelon, Brian and A. James O'Malley. "Bayesian Analysis Using Power Priors with Application to Pediatric Quality of Care". In: *Journal of Biometrics & Biostatistics* 01.01 (2010), pp. 1–9. DOI: 10.4172/2155-6180.1000103.

Psioda, Matthew A. and Joseph G. Ibrahim. "Bayesian clinical trial design using historical data that inform the treatment effect". In: *Biostatistics* 20.3 (2019), pp. 400–415. ISSN: 14684357. DOI: 10.1093/biostatistics/kxy009.

Rosenberg, David S. *Completing the Square*. Tech. rep. URL: moz-extension://e82277a5-b6c9-ac4b-adfe-5ff0f041bf55/enhanced-reader.html?openApp%7B%5C&%7Dpdf=https%7B%5C%%7D3A%7B%5C%%7D2F%7B%5C%%7D2Fdavidrosenberg.github.io%7B%5C%%7D2Fmlcourse%7B%5C%%7D2FNotes%7B%5C%%7D2Fcompleting-the-square.pdf.

Saville, Benjamin R., Jason T. Connor, Gregory D. Ayers, and Joann Alvarez. "The utility of Bayesian predictive probabilities for interim monitoring of clinical trials". In: *Clinical Trials* 11.4 (2014), pp. 485–493. ISSN: 17407753. DOI: 10.1177/1740774514531352.

Sieck, Victoria R.C. and Fletcher G.W. Christensen. "A framework for improving the efficiency of operational testing through Bayesian adaptive design". In: *Quality and Reliability Engineering International* (2020), pp. 1–16. ISSN: 10991638. DOI: 10.1002/qre.2802.

Test and Evaluation Management Guide. Fort Belvoir, VA: The Defense Acquisition University Press, 2005. URL: http://www.acqnotes.com/Attachments/Test%20and%20Evaluation%20Management%20Guide.pdf.

Thall, Peter F., Leiko H. Wooten, Christopher J. Logothetis, Randall E. Millikan, and Nizar M. Tannir. "Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censorin". In: 26 (2007), pp. 4687–4702. DOI: 10.1002/sim.2894.

Yin, Guosheng, Nan Chen, and J. Jack Lee. "Phase II trial design with Bayesian adaptive randomization and predictive probability". In: *Journal of the Royal*

Statistical Society. Series C: Applied Statistics 61.2 (2012), pp. 219–235. ISSN: 00359254. DOI: 10.1111/j.1467-9876.2011.01006.x.

Zang, Yong and J. Jack Lee. "Adaptive clinical trial designs in oncology". In: *Chinese Clinical Oncology* 3.4 (2014). ISSN: 23043873. DOI: 10.3978/j.issn.2304-3865.2014.06.04.

Zhou, Ming, Qi Tang, Lixin Lang, Jun Xing, and Kay Tatsuoka. "Predictive probability methods for interim monitoring in clinical trials with longitudinal outcomes". In: *Statistics in Medicine* 37.14 (2018), pp. 2187–2207. ISSN: 10970258. DOI: 10.1002/sim.7685.